Real-Time Sound Source Localization in Videoconferencing Environments

Author: Martí Guerola, Amparo

Directors: Cobos Serrano, Máximo
López Monfort, José Javier

Abstract

Sound Source Localization (SSL) mechanisms have been extensively studied. Many applications like teleconferencing or speech enhancement systems require the localization of one or more acoustic sources. Moreover, it is essential to localize sources also in noisy and reverberant environments. It has been shown that computing the Steered Response Power (SRP) is more robust approach than two-stage, direct time-difference of arrival methods. The problem with computing the SRP is that a fine grid search procedure is needed, which is too expensive for a real-time system. To this end, it has been introduced a new strategy (modified SRP-PHAT functional) which can be used for a real-time system with a low computational cost. Moreover, it has been demonstrated that the statistical distribution of location estimates when a speaker is active can be successfully used to discriminate between speech and non-speech frames. The main objective of this work is to describe our new localization approach and integrate it into a real-time speaker localization and detection system. The applicability of the method will be shown for a real videoconferencing environment using an acoustically-driven steering camera.

Resumen

Los mecanismos de Localización de Fuentes de Sonido (SSL) han sido ampliamente estudiados. Muchas aplicaciones como sistemas de teleconferecia o realzado de voz necesitan la localización de una o más fuentes acústicas. Además es esencial localizar las fuentes incluso en ambientes ruidosos y con reverberación. Se ha demostrado que el Steered Response Power (SRP) es un método más robusto que los métodos de dos pasos basados en la diferencia de tiempo de llegada. El problema en el cálculo del SRP es que es necesario el uso de un mallado fino lo que implica un coste computacional muy alto para ser utilizado en sistemas de tiempo real. Con este propósito, se ha introducido una nueva estrategia (función modificada SRP-PHAT) que puede ser usada en un sistema de tiempo real con un coste computacional bajo. Además se ha demostrado que la distribución estadística de las posiciones estimadas cuando el hablante está activo puede ser utilizado satisfactoriamente para distinguir fragmentos de habla y no habla. El principal objetivo de este trabajo es describir nuestra nueva propuesta e integrarla en un sistema de localización y detección de hablantes en tiempo real. Se mostrará la aplicabilidad del método en un entorno real de videoconferencia usando una cámara acústicamente dirigida.

Author: Martí Guerola, Amparo, email: ammargue@upvnet.upv.es
Directors: Cobos Serrano, Máximo, email: macoser1@iteam.upv.es
López Monfort, José Javier, email: jjlopez@dcom.upv.es
Submitting Date: 26-11-10

Contents

Chapter 1

Introduction

The localization of sources of emitting signals has been the focus of attention for more than a century. Localization and aiming in addition to noise and interference rejection allow microphone arrays to outperform single microphone systems. Arrays of microphones have a variety of applications in speech data acquisition systems. Applications include teleconferencing, biomedical devices for hearing impaired persons, audio surveillance, gunshot detection and camera pointing systems. The fundamental requirement for sensor array systems is the ability to locate and track a signal source. In addition to having high accuracy, the location estimator must be capable of a high update rate at reasonable computational cost in order to be useful for real time tracking and beamforming applications. Source location data may also be used for purposes other than beamforming, such as aiming a camera in a video conferencing system (Fig.1.1).

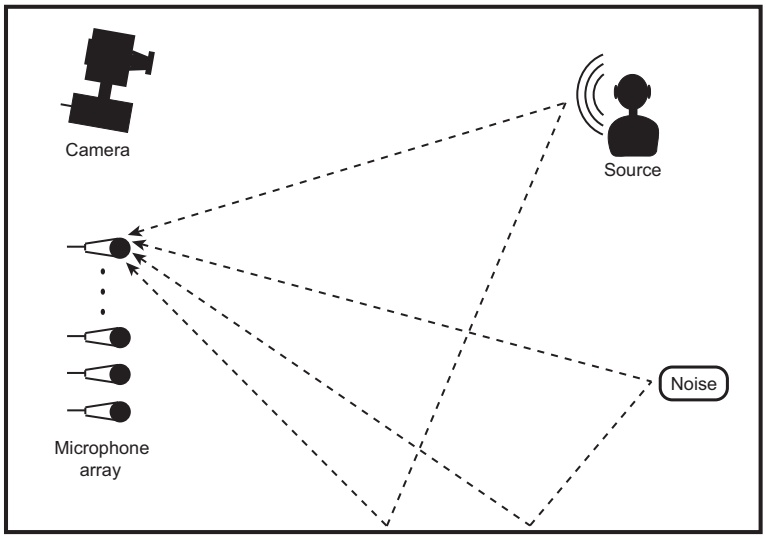


Figure 1.1: Sound source localization problem in an enclosed area.

Many current SSL systems assume that the sound sources are distributed on a horizontal plane [2]. This assumption simplifies the problem of SSL in almost all methods. In teleconference ap-

plications they assume all talkers speak at the same height which is somewhat true, but the talker or other attendees can act as sound blockades between the main talker and the array, which is typically a linear wall-mounted microphone array. Moreover, in most dominant SSL methods, the computational cost for two dimensional cases is high so that the real time implementation needs a computer with high processing power. Some of these SSL methods have been modified to cover a three dimensional space at a very high computational cost. There is thus a need for a SSL technique in 3D space that can be implemented in real time without requiring high computational power. There is also another problem to take into account, that is the reflections of the sound signal in the different walls, floor or objects which there are around. These reflections interfere in the system making more difficult the localization so then, the SSL systems must be robust and work in adverse conditions: noisy and reverberant environments (see Fig.1.1).

Figure 1.2: Videoconferencing room (Cisco Telepresence 3000).

In this work we propose to use an improved SSL technique to develop an automatic voice-steering camera application. The objective is to be able to localize the members of a videoconferece taking place in a room. To this aim the implemented algorithm must be able to work in real time so its computational cost can not be as high as the conventional SSL algorithms. Moreover, taking into account that the speakers are quite close each one to the others, the technique employed must be robust and precise enough to identify correctly the main speaker. Once the speaker is located, the coordinates of his/her position are sent to a camera which points to the face of the member who is talking in this moment, making the video conference more similar to face to face communication (see Fig.1.2).

This Master's thesis is organized as follows. In Chapter 2, conventional SSL techniques are reviewed. Chapter 3 discusses the advantages of a modified SSL algorithm proposed by the author.

This approach is compared to other SSL methods in Chapter 4. Chapter 5 introduce a speaker detection based on the statistics of the resulting location estimates by the proposed SSL algorithm. The application of this approach and the speaker detector to videoconferencing environment is shown in Chapter 6. Finally, the conclusions obtained from all these experiments are presented in Chapter 7.

Chapter 2

Sound Source Localization

Sound Source Localization (SSL) is the process of determining the spatial location of a sound source based on multiple observations of the emitted sound signal [15]. The existing strategies of SSL may broadly be divided into two main classes: indirect and direct approaches [14]. Indirect approaches to source localization are usually two-step methods: first, the relative time delays for the various microphone pairs are evaluated and then the source location is found as the intersection of a pair of a set of half-hyperboloids centered around the different microphone pairs. Each half-hyperboloid determines the possible location of a sound source based on the measure of the time difference of arrival between the two microphones. On the other hand, direct approaches generally scan a set of candidate source positions and pick the most likely candidate as an estimate of the sound source location, thus performing the localization in a single step.

For both approaches, techniques such as the Generalized Cross-Correlation (GCC) method, proposed by Knapp and Carter in 1976, are widely used [13]. The Time Delay Estimation (TDE) between signals from any pair of microphones can be performed by computing the cross-correlation function of the two signals after applying a suitable weighting step. The lag at which the cross-correlation function has its maximum is taken as the time delay between them.

The type of weighting used with GCC is crucial to localization performance. Among several types of weighting, the phase transform (PHAT) is the most commonly used pre-filter for the GCC because it is more robust against reverberation. The GCC with the phase transform (GCC-PHAT) approach has been shown to perform well in a mild reverberant environment. Unfortunately, in the presence of even moderate reverberation levels, the algorithm is seriously hampered, due to the presence of spurious peaks. Also reflections of the signal on the walls make appear different peaks in the impulse response of the room which can generate peaks in the GCC function that may be strongest than the peak corresponding to the direct path. An example room impulse response is shown Figure 2.1.

Another class of important SSL algorithms is that based on a steered beamformer. When the source location is not known, a beamformer can be used to scan over a predefined spatial region by adjusting its steering parameters. The output of a beamformer is known as the steered response. When the point or direction of scan matches the source location, the SRP will be maximized. How-

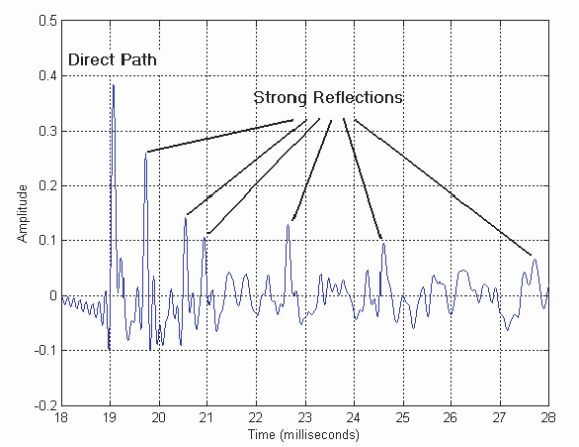


Figure 2.1: Room impulse response from source to one microphone.

ever, the localization performance of the conventional steered-beamformer techniques which apply filters to the array signals have been derived to improve its performance. When the phase transform filter is incorporated with the steered-beamformer method, the resulting algorithm (SRP-PHAT) is superior in combating the adverse effects of background noise and reverberation compared to the conventional steered-beamformer method and the pairwise method, GCC-PHAT [13].

In the present day, the SRP-PHAT algorithm has become the most popular localization method for its good robust performance in real environment. However, the computational requirements of the method are large and this makes real-time implementation difficult. Since the SRP-PHAT method was proposed, there have been several attempts to reduce the computational requirements of the intrinsic SRP search process [16],[4].

Other approaches to sound localization include Multiple Signal Classification (MUSIC) [11], [23], and Maximum Likelihood (ML) estimation [24], though these are typically applied to far-field narrow band direction-of- arrival estimation problems [15].

In the next subsections we introduce the concept of time delay estimation which is necessary for the SSL task. Then, the SRP is deeply explained when using the phase transform pre-filter.

2.1 Time Difference Of Arrival (TDOA)

Most practical acoustic source localization schemes are based on *Time Delay Of Arrival* estimation (TDOA) for the following reason: such systems are conceptually simple. They are reasonably effective in moderately reverberant environments and, moreover, their low computational complexity makes them well-suited to real-time implementation with several sensors [21].

In general, an array is composed of M microphones, and each microphone is positioned at a unique spatial location. Hence, the direct-path sound waves propagate along M bearing lines,

from the source to each microphone, simultaneously. The orientations of these lines in the global coordinate system define the propagation directions of the wave fronts at each microphone. The propagation vectors for a four-element ($m = 1, \ldots, 4$), linear array are illustrated in Figure 2.2, denoted as $\vec{d}_m$.

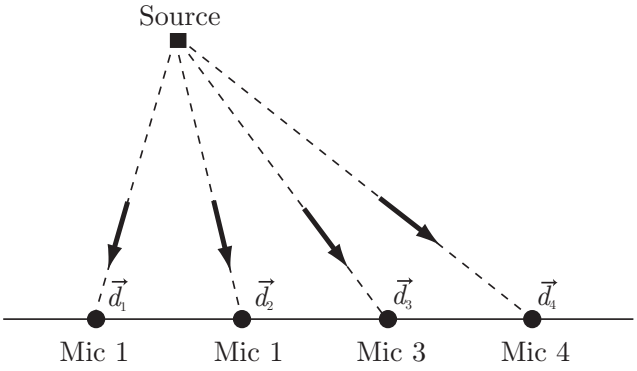


Figure 2.2: Propagation vectors.

Time Delay Estimation (TDE) is concerned with the computation of the relative TDOA between different microphone sensors. It is a fundamental technique in microphone array signal processing and the first step in passive TDOA-based acoustic source localization systems. With this kind of localization, a two-step strategy is adopted as shown in Fig. 2.3.

The first stage involves estimation of the TDOA between receivers through the use of TDE techniques [5]. The estimated TDOAs are then transformed into range difference measurements between sensors, resulting in a set of nonlinear hyperbolic range difference equations. The second stage utilizes efficient algorithms to produce an unambiguous solution to these nonlinear hyperbolic equations. The solution produced by these algorithms result in the estimated position location of the source [19]. This data along with knowledge of the microphone positions are then used to generate hyperbolic curves, which are then intersected in some optimal sense to arrive at a source location estimate as shown in Figure 2.4.

Several variations of this principle have been developed [20]. They differ considerably in the method of derivation, the extent of their applicability (2D versus 3D, near field source versus far field source), and their means of solution.

2.2 SRP using the Phase Transform (PHAT)

Array signal processing techniques rely on the ability to focus on signals originating from a particular location or direction in space. Most of these techniques employ some type of beamforming, which generally includes any algorithm that exploits an array's sound-capture ability [12]. Beamforming, in the conventional sense, can be defined by a filter-and-sum process, which applies some

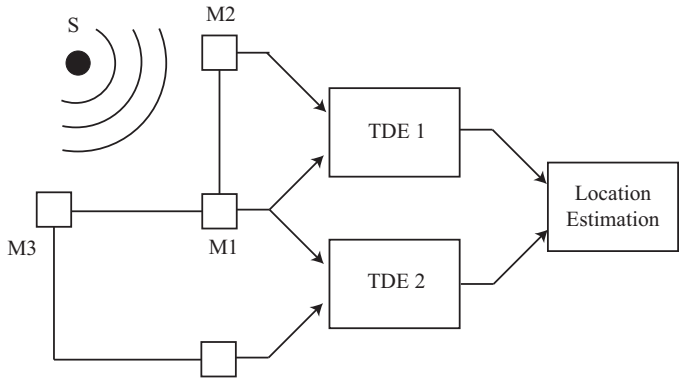


Figure 2.3: A two stage algorithm for sound source localization.

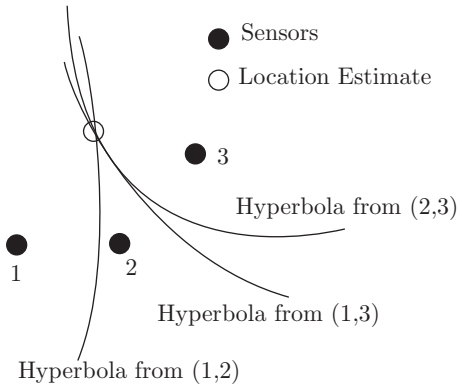


Figure 2.4: Source estimation with three microphones.

temporal filters to the microphone signals before summing them to produce a single, focused signal. These filters are often adapted during the beamforming process to enhance the desired source signal while attenuating others. The simplest filters execute time shifts that have been matched to the source signals propagation delays. This method is referred to as delay-and-sum beamforming; it delays the microphone signals so that all versions of the source signal are time-aligned before they are summed. The filters of more sophisticated filter-and-sum techniques usually apply this time alignment as well as other signal-enhancing processes.

Beamforming techniques have been applied to both source-signal capture and source localization. If the location of the source is known (and perhaps something about the nature of the source signal is known as well), then a beamformer can be focused on the source, and its output becomes an enhanced version (in some sense) of the inputs from the microphones. If the location of the source is not known, then a beamformer can be used to scan, or steer, over a predefined spatial region by adjusting its steering delays (and possibly its filters). As previously commented, the output of a beamformer, when used in this way, is known as the steered response. The SRP may peak under a variety of circumstances, but with favorable conditions, it is maximized when

the steering delays match the propagation delays. By predicting the properties of the propagating waves, these steering delays can be mapped to a location, which should coincide with the location of the source.

For voice capture application, the filters applied by the filter-and-sum technique must not only suppress the background noise and contributions from unwanted sources, they must also do this in way that does not significantly distort the desired signal. The most common of these filters is the phase transform (PHAT), which applies a magnitude-normalizing weighting function to the cross-spectrum of two microphone signals.

We now describe the measurement principle of SRP-PHAT algorithm which is closely related to GCC-PHAT, and then introduce its implementation.

2.2.1 SRP-PHAT algorithm

Consider the output from microphone l, $m_l(t)$, in an M microphone system. Then, the SRP at the spatial point $\mathbf{x} = [x, y, z]$ for a time frame n of length T is defined as

$$P_n(\mathbf{x}) \equiv \int_{nT}^{(n+1)T} \left| \sum_{l=1}^{M} w_l m_l \left(t - \tau(\mathbf{x}, l)\right) \right|^2 dt, \tag{2.1}$$

where w_l is a weight and $\tau(\mathbf{x}, l)$ is the direct time of travel from location $\mathbf{x}$ to microphone l.

DiBiase [7] showed that the SRP can be computed by summing the GCCs for all possible pairs of the set of microphones. The GCC for a microphone pair (k, l) is computed as

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \Phi_{kl}(\omega) M_k(\omega) M_l^*(\omega) e^{j\omega\tau} d\omega, \tag{2.2}$$

where τ is the time lag, * denotes complex conjugation, $M_l(\omega)$ is the Fourier transform of the microphone signal $m_l(t)$, and $\Phi_{kl}(\omega) = W_k(\omega) W_l^*(\omega)$ is a combined weighting function in the frequency domain. The Phase Transform (PHAT) [13] has been demonstrated to be a very effective GCC weighting for time delay estimation in reverberant environments:

$$\Phi_{kl}(\omega) \equiv \frac{1}{|M_k(\omega) M_l^*(\omega)|}. \tag{2.3}$$

Taking into account the symmetries involved in the computation of Eq.(2.1) and removing some fixed energy terms [7], the part of $P_n(\mathbf{x})$ that changes with $\mathbf{x}$ is isolated as

$$P'_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} R_{m_k m_l}\left(\tau_{kl}(\mathbf{x})\right), \tag{2.4}$$

where $\tau_{kl}(\mathbf{x})$ is the *Inter-Microphone Time-Delay Function* (IMTDF). This function is very important, since it represents the theoretical direct path delay for the microphone pair (k, l) resulting from a point source located at $\mathbf{x}$. The IMTDF is mathematically expressed as

$$\tau_{kl}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{x}_k\| - \|\mathbf{x} - \mathbf{x}_l\|}{c}, \tag{2.5}$$

where c is the speed of sound, and $\mathbf{x}_k$ and $\mathbf{x}_l$ are the microphone locations.

The SRP-PHAT algorithm consists in evaluating the functional $P'_n(\mathbf{x})$ on a fine grid G with the aim of finding the point-source location $\mathbf{x}_s$ that provides the maximum value:

$$\mathbf{x}_s = \arg\max_{\mathbf{x}\in G} P'_n(\mathbf{x}). \tag{2.6}$$

2.2.2 Implementation

Basically, the SRP-PHAT algorithm is implemented as follows:

- Define a spatial grid G with a given spatial resolution r. The theoretical delays from each point of the grid to each microphone pair are pre-computed using Eq.(2.5).
- For each analysis frame, the GCC of each microphone pair is computed as expressed in Eq.(2.2).
- For each position of the grid $\mathbf{x} \in G$, the contribution of the different cross-correlations are accumulated (using delays pre-computed in 1), as in Eq.(2.4).
- Finally, the position with the maximum score is selected.

2.2.3 Other modifications

The accuracy of the SRP-PHAT algorithm is limited by the time resolution of the PHAT weighted cross correlation functions [22]. However, despite its robustness, computational cost is a real issue because the SRP space to be searched has many local extrema [1]. Very interesting modifications have already been proposed to improve the SRP-PHAT algorithm. Some of this modifications only affect to the weighting factor. In [17] until five different weighting factors are proposed to improve the precision of the localization. Also exists the PHAT-β transform which varies the degree of spectral magnitude information (partial whitening) of each microphone signal using a single parameter, β, which varies from 0 (no whitening) to 1 (total whitening). A simulation study described in [10] considered the detection performance of sound sources using the PHAT-β and they have demonstrated that the standard PHAT (β = 1) improves detection performance for broadband signals. However, the optimal choice of β typically ranged between 0.5 and 0.8, which resulted in a significant performance improvements over both total (β=1) and no whitening (β=0).

$$\Phi_{kl}(\omega) \equiv \frac{1}{|M_k(\omega)M_l^*(\omega)|^\beta}. \tag{2.7}$$

While many transforms consider improving SNR, the PHAT-β primarily deconvolves the spectrum so that each frequency region contributes more uniformly to the coherent sum of the steered power.

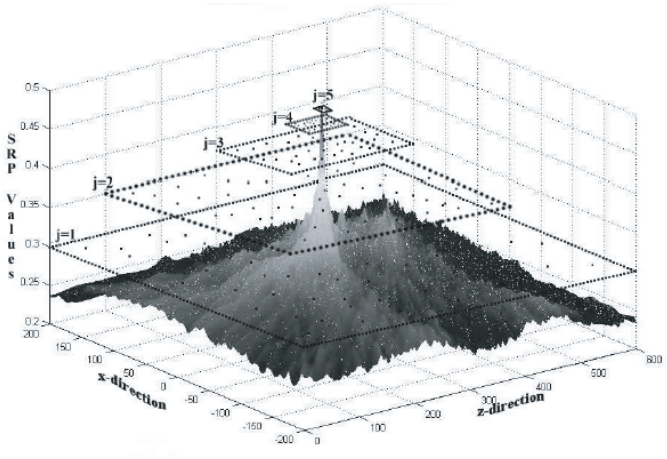


Figure 2.5: 2D example of SRC: j is the iteration index. The rectangular regions show the contracting search regions.

Other modifications of the SRP-PHAT algorithm are focused on reducing the computational cost of that technique. Examples of them are those based on Stochastic Region Contraction (SRC) [7] and Coarse-to-Fine Region Contraction (CFRC) [9]. The first proposes, using SRC, to make computing the SRP practical. So it is given an initial rectangular search volume containing the desired global optimum and perhaps many local maxima or minima, gradually, in an iterative process, contract the original volume until a sufficiently small subvolume is reached in which the global optimum is trapped (see Fig. 2.5). The second proposal uses a CFRC to make computing the SRP practical as well. Using CFRC can reduce the computational cost by more than three orders of magnitude [8].

Chapter 3

Improved SRP-PHAT algorithm for Source Localization

A different strategy for implementing a less cost computational SRP-PHAT algorithm is shown in this section. The algorithm proposed instead of evaluating the SRP functional at discrete positions of a spatial grid, it is integrated over the GCC lag space corresponding to the volume surrounding each point of the grid [6].

3.1 The Inter-Microphone Time Delay Function

As commented in the previous chapter, the IMTDF plays a very important role in the source localization task. This function can be interpreted as the spatial distribution of possible TDOAs resulting from a given microphone pair geometry.

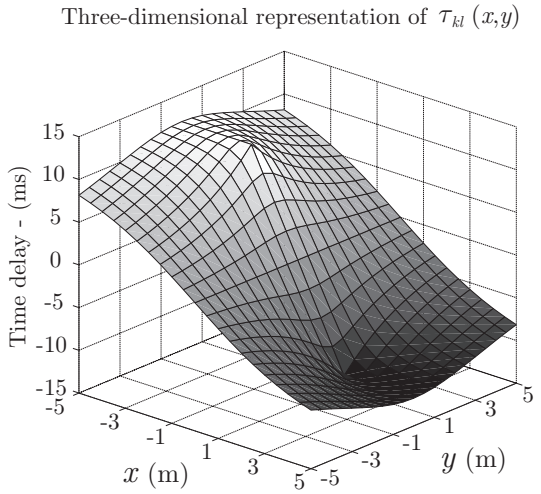


Figure 3.1: Example of IMTDF. Representation for the plane $z = 0$ with microphones located at $[-2, 0, 0]$ and $[2, 0, 0]$.

The function $\tau_{kl}(\mathbf{x})$ is continuous in $\mathbf{x}$ and changes rapidly at points close to the line connecting both microphone locations. Therefore, a pair of microphones used as a time-delay sensor is maximally sensible to changes produced over this line. An example function is depicted in Figure 3.1 for the plane $z = 0$, with $\mathbf{x}_k = [-2, 0, 0]$ and $\mathbf{x}_l = [2, 0, 0]$. The gradient of the function is shown in Figure 3.2.

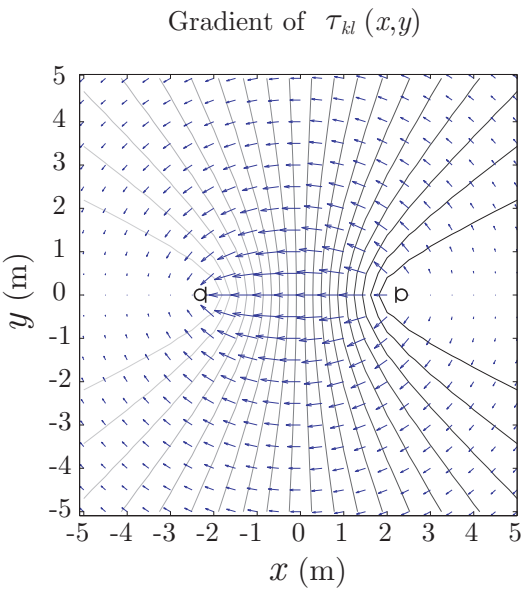


Figure 3.2: Example of IMTDF. Gradient.

It is useful here to remark that the equation $|\tau_{kl}(\mathbf{x})| = C$, with C being a positive real constant, defines a hyperboloid in space with foci on the microphone locations $\mathbf{x}_k$ and $\mathbf{x}_l$. Moreover, the set of continuous confocal half-hyperboloids $\tau_{kl}(\mathbf{x}) = C$ with $C \in [-C_{\max}, C_{\max}]$, being $C_{\max} = (1/c)\|\mathbf{x}_k - \mathbf{x}_l\|$, spans the whole three-dimensional space.

At this point we can formulate the next theorem: Given a volume V in space, the IMTDF for points inside V, $\tau_{kl}(\mathbf{x} \in V)$, takes only values in the continuous range $[\min(\tau_{kl}(\mathbf{x} \in \partial V)), \max(\tau_{kl}(\mathbf{x} \in \partial V))]$, where ∂V is the boundary surface that encloses V.

In order to prove the theorem above, let us assume that a point inside V, $\mathbf{x}_0 \in V$, takes the maximum value in the volume, i.e. $\tau_{kl}(\mathbf{x}_0) = \max(\tau_{kl}(\mathbf{x} \in V)) = C_{\max_V}$. Since there is a half-hyperboloid that goes through each point of the space, all the points besides $\mathbf{x}_0$ satisfying $\tau_{kl}(\mathbf{x}) = C_{\max_V}$ will also take the maximum value. Therefore, all the points on the surface resulting from the intersection of the volume and the half-hyperboloid will take this maximum value, including those pertaining to the boundary surface ∂V. The existence of the minimum in ∂V is similarly deduced.

The above property is very useful to understand the advantages of the approach presented in this work. Note that the SRP-PHAT algorithm is based on accumulating the values of the different GCCs at those time lags coinciding with the theoretical inter-microphone time delays, which are

only computed at discrete points of a spatial grid. However, as described before, it is possible to analyze a complete spatial volume by scanning the time-delays contained in a range defined by the maximum and minimum values on its boundary surface. In the section 3.2, we describe how this knowledge can be included in the localization algorithm to increase its robustness.

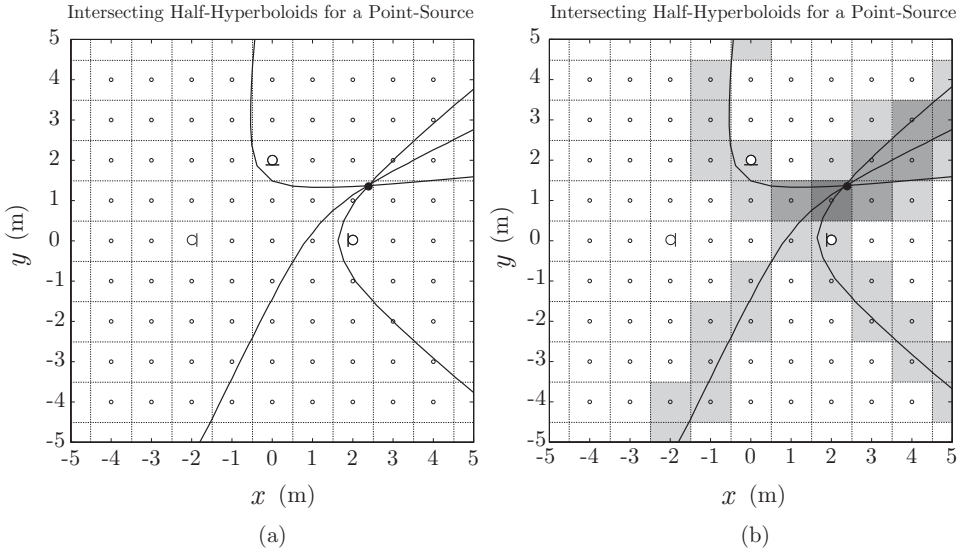


Figure 3.3: Intersecting half-hyperboloids and localization approaches. (a) Conventional SRP-PHAT. (b) Proposed.

3.2 Proposed Approach

Let us begin the description of the proposed approach by analyzing a simple case where we want to estimate the location $\mathbf{x}_s$ of a sound source inside an anechoic space. In this simple case, the GCCs corresponding to each microphone pair are delta functions centered at the corresponding inter-microphone time-delays: $R_{m_k m_l}(\tau) = \delta(\tau - \tau_{kl}(\mathbf{x}_s))$. For example and without loss of generality, let us assume a set-up with $M = 3$ microphones, as depicted in Figure 3.3(a). Then, the source position would be that of the intersection of the three half-hyperboloids $\tau_{kl}(\mathbf{x}) = \tau_{kl}(\mathbf{x}_s)$, with $(k, l) \in \{(1, 2), (1, 3), (2, 3)\}$. Consider now that, to localize the source, a spatial grid with resolution $r = 1$ m is used as shown in Figure 3.3(a). Unfortunately, the intersection does not coincide with any of the sampled positions, leading to an error in the localization task. Obviously, this problem would have been easier to solve with a two step localization approach, but the above example shows the limitations imposed by the selected spatial sampling in SRP-PHAT, even in optimal acoustic conditions. This is not the case of the approach followed to localize the source in Figure 3.3(b) where, using the same spatial grid, the GCCs have been integrated for each sampled position in a range that covers their volume of influence. A darker gray color indicates a greater accumulated value and, therefore, the darkest area is being correctly identified as the one containing the true sound source location. This new modified functional is expressed as follows

$$P_n''(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \sum_{\tau=L_{kl1}(\mathbf{x})}^{L_{kl2}(\mathbf{x})} R_{m_k m_l}(\tau). \tag{3.1}$$

The problem is to determine correctly the limits $L_{kl1}(\mathbf{x})$ and $L_{kl2}(\mathbf{x})$, which depend on the specific IMTDF resulting from each microphone pair. The computation of these limits is explained in the next subsection.

3.2.1 Computation of integration limits

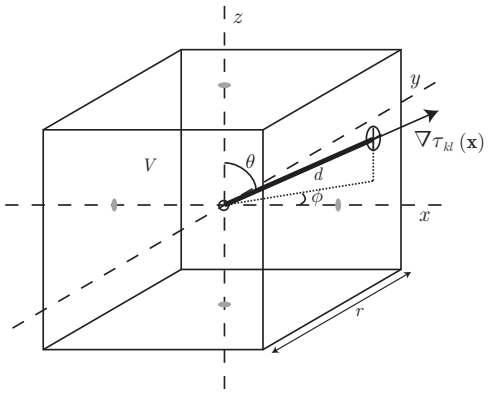


Figure 3.4: Volume of influence of a point in a rectangular grid.

As explained in Section 3.1, the IMTDF inside a volume can only take values in the range defined by its boundary surface. Therefore, for each point of the grid, the problem of finding the GCC integration limits of its volume of influence can be simplified to finding the maximum and minimum values on the boundary. To this end, it becomes useful to study the direction of the greatest rate of increase at each grid point, which is given by the gradient

$$\nabla\tau_{kl}(\mathbf{x}) = \left[\nabla_x\tau_{kl}(\mathbf{x}), \nabla_y\tau_{kl}(\mathbf{x}), \nabla_z\tau_{kl}(\mathbf{x})\right], \tag{3.2}$$

where

$$\begin{aligned}
\nabla_x\tau_{kl}(\mathbf{x}) = \frac{\partial\tau_{kl}(\mathbf{x})}{\partial x} &= \frac{1}{c}\left(\frac{x-x_k}{\|\mathbf{x}-\mathbf{x}_k\|} - \frac{x-x_l}{\|\mathbf{x}-\mathbf{x}_l\|}\right), \\
\nabla_y\tau_{kl}(\mathbf{x}) = \frac{\partial\tau_{kl}(\mathbf{x})}{\partial y} &= \frac{1}{c}\left(\frac{y-y_k}{\|\mathbf{x}-\mathbf{x}_k\|} - \frac{y-y_l}{\|\mathbf{x}-\mathbf{x}_l\|}\right), \\
\nabla_z\tau_{kl}(\mathbf{x}) = \frac{\partial\tau_{kl}(\mathbf{x})}{\partial z} &= \frac{1}{c}\left(\frac{z-z_k}{\|\mathbf{x}-\mathbf{x}_k\|} - \frac{z-z_l}{\|\mathbf{x}-\mathbf{x}_l\|}\right).
\end{aligned} \tag{3.3}$$

The integration limits can be calculated for a symmetric volume by taking the product of the magnitude of the gradient and the distance d that exists from the grid point to the intersection of

a line with the gradient's direction and the boundary:

$$L_{kl1}(\mathbf{x}) = \tau_{kl}(\mathbf{x}) - \|\nabla\tau_{kl}(\mathbf{x})\| \cdot d, \tag{3.4}$$

$$L_{kl2}(\mathbf{x}) = \tau_{kl}(\mathbf{x})) + \|\nabla\tau_{kl}(\mathbf{x})\| \cdot d, \tag{3.5}$$

Figure 3.4 depicts the geometry for a rectangular grid with spatial resolution r. For this cubic geometry, the distance d can be expressed as

$$d = \frac{r}{2} \min\left(\frac{1}{|\sin\theta\cos\phi|}, \frac{1}{|\sin\theta\sin\phi|}, \frac{1}{|\cos\theta|}\right), \tag{3.6}$$

where

$$\theta = \cos^{-1}\left(\frac{\nabla_z\tau_{kl}(\mathbf{x})}{\|\nabla\tau_{kl}(\mathbf{x})\|}\right), \tag{3.7}$$

$$\phi = \text{atan}_2(\nabla_y\tau_{kl}(\mathbf{x}), \nabla_x\tau_{kl}(\mathbf{x})), \tag{3.8}$$

being $\text{atan}_2(y, x)$ the quadrant-resolving arctangent function.

3.2.2 Computational Cost

Let L be the DFT length of a frame and $Q = M(M-1)/2$ the number of microphone pairs. The computational cost of SRP-PHAT is given by [18]:

$$\begin{aligned}\text{SRP-PHAT}_{cost} \approx\ & [6.125Q^2 + 3.75Q]L\log_2 L \\ & +15LQ(1.5Q-1) + (45Q^2 - 30Q)\nu',\end{aligned} \tag{3.9}$$

where ν' is the average number of functional evaluations required to find the maximum of the SRP space. Since the cost added by the modified functional is negligible and the frequency-domain processing of our approach remains the same as the conventional SRP-PHAT algorithm, the above formula is valid for both approaches. Moreover, since the integration limits can be pre-computed before running the localization algorithm, the associated processing does not involve additional computation effort. However the advantage of the proposed method relies on the reduced number of required functional evaluations ν' for detecting the true source location, which results in an improved computational efficiency.

Chapter 4

SSL Comparative

First of all it was necessary to demonstrate that the modified SRP-PHAT algorithm proposed has a similar behavior to traditional SRP-PHAT, so different experiments have been carried out.

4.1 Description of the application

Different experiments with real and synthetic recordings were conducted to compare the performances of the conventional SRP-PHAT algorithm, the SRC algorithm (explained in 2.2.3) and our proposed method. First, the *Roomsim* Matlab package [3] was used to simulate an array of 6 microphones placed on the walls of a shoe-box-shaped room with dimensions 4 m $\times$ 6 m $\times$ 2 m (Fig. 4.1).

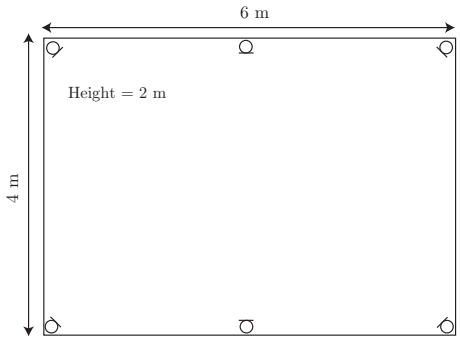


Figure 4.1: Set-up.

4.2 Results

The simulations were repeated with two different reverberation times ($T_{60} = 0.2$s and $T_{60} = 0.7$s), considering 30 random source locations and different Signal-to-Noise Ratio (SNR) conditions. The resultant recordings were processed with 3 different spatial grid resolutions in the case of SRP-PHAT and the proposed method ($r_1 = 0.01$ m, $r_2 = 0.1$ m and $r_3 = 0.5$ m). Note that the number of

functional evaluations ν' depends on the selected value of r, having $\nu'_1 = 480 \times 10^5$, $\nu'_2 = 480 \times 10^2$ and $\nu'_3 = 384$. The implementation of SRC was the one made available by Brown University's LEMS at `http://www.lems.brown.edu/array/download.html`, using 3000 initial random points. The processing was carried out using a sampling rate of 44.1 kHz, with time windows of 4096 samples of length and 50% overlap. The simulated sources were male and female speech signals of length 5 s with no pauses. The averaged results in terms of *Root Mean Squared Error* (RMSE) are shown in Figure 4.2(a-c).

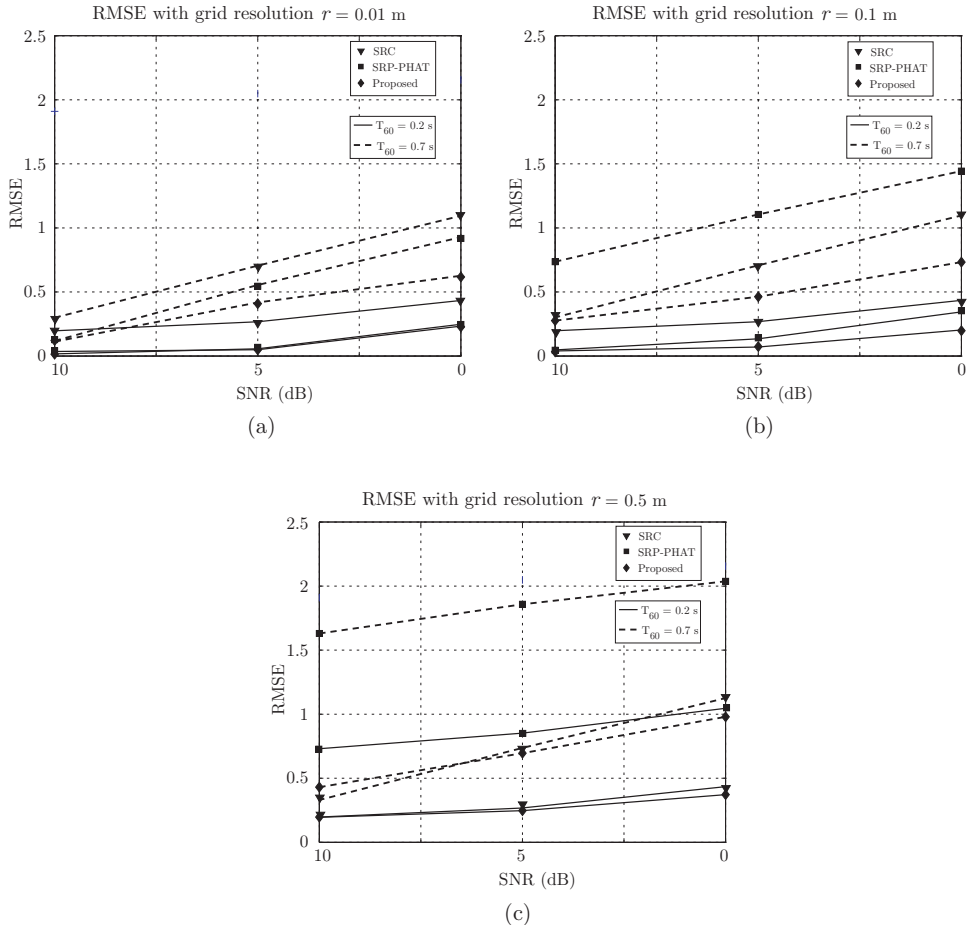
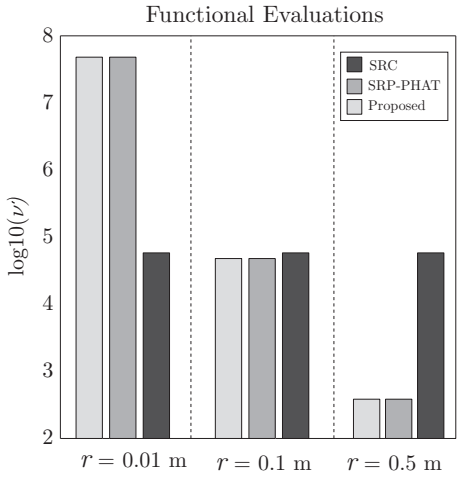


Figure 4.2: Results with simulations. (a) $r = 0.01$ m. (b) $r = 0.1$ m. (c) $r = 0.5$ m.

Since SRC does not depend on the grid size, the SRC curves are the same in all these graphs. As expected, all the tested systems perform considerably better in the case of low reverberation and high SNR. For the finest grid, it can be clearly observed that the performance of SRP-PHAT and the proposed method is almost the same. However, for coarser grids, our proposed method is only slightly degraded, while the performance of SRP-PHAT becomes substantially worse, specially for low SNRs and high reverberation. SRC has similar performance to SRP-PHAT with $r = 0.01$ m. Therefore, our proposed approach performs robustly with higher grid sizes, which results in a great computational saving in terms of functional evaluations, as depicted in Figure 4.3.



(e)

Figure 4.3: Functional evaluations.

r	0.01	0.1	0.5
ν'	$802 \cdot 10^5$	$802 \cdot 10^2$	641
SRP-PHAT	RMSE = 0.29	RMSE = 0.74	RMSE = 1.82
Proposed	RMSE = 0.21	RMSE = 0.29	RMSE = 0.31
SRC	RMSE = 0.34 ($\nu' = 58307$)		

Table 4.1: RMSE for the real-data experiment.

On the other hand, a real set-up quite similar to the simulated one was considered to study the performance of the method in a real scenario. Six omnidirectional microphones were placed at the 4 corners and at the middle of the longest walls of a videoconferencing room with dimensions 5.7 m $\times$ 6.7 m $\times$ 2.1 m and 12 seats. The measured reverberation time was $T_{60} = 0.28$ s. The processing was the same as with the synthetic recordings, using continuous speech fragments obtained from the 12 seat locations. The results are shown in Table 4.1 and confirm that our proposed method performs robustly using a very coarse grid.

Although similar accuracy to SRC is obtained, the number of functional evaluations is significantly reduced.

Figure 4.4 shows that, for a fine grid, there is no difference between traditional and modified SRP-PHAT method. Note that the GCC resulting from each pair of microphones cross in the same point with equal accuracy. However, figures (a) and (b) of Fig.4.5 show that the results of localization when a coarse grid is used in the GCC calculations have not equal accuracy if traditional or modified SRP-PHAT is applied. It can be seen that when a coarse grid is used in order to get lower computational cost, the traditional SRP-PHAT approach has not enough accuracy to find the SSL while the proposed modified SRP-PHAT is precise enough.

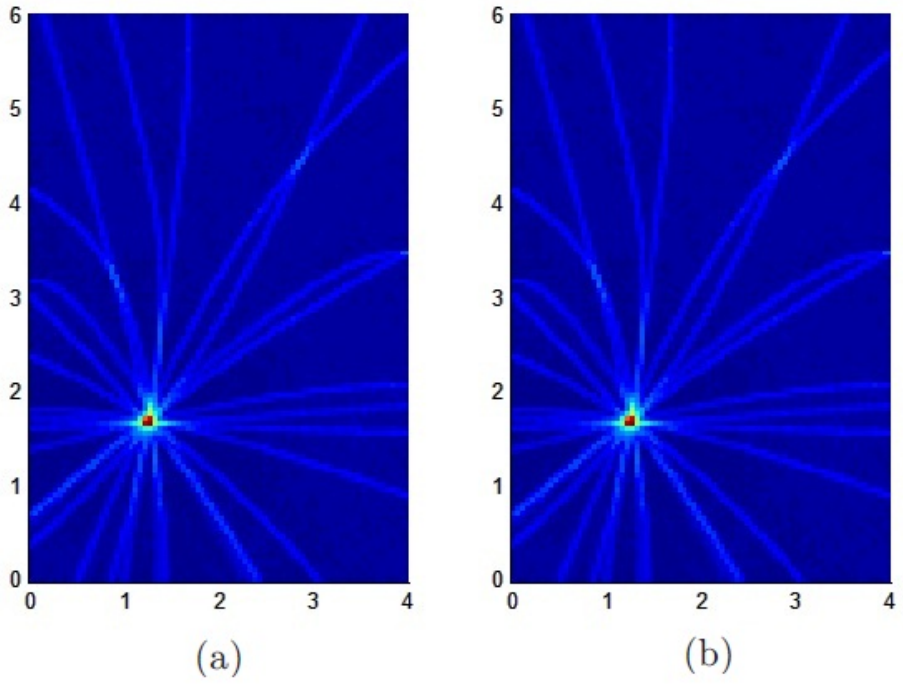


Figure 4.4: Source likelihood map. Fine grid (a) traditional and (b) modified SRP-PHAT.

Another way to evaluate the benefits of our proposed approach is by looking at the results shown in Table 4.2. It shows the percentage of correct frames were the source was correctly located using our proposed approach and the conventional SRP-PHAT algorithm. A frame estimate is considered to be erroneous if its deviation from the true source location is higher than 0.4 m, which is approximately the maximum deviation admissible for the coarser grid. Notice that, for the worst case ($T_{60} = 0.7$ and SNR$= 0$ dB), the proposed approach is capable of localizing correctly the source with 74% correctness with $r = 0.5$ m, which is approximately the performance achieved by the conventional algorithm using $r = 0.01$ m. Thus, our proposed approach provides similar performance with a reduction of five orders of magnitude in the required number of functional

Method (T_{60})	Source 1 SNR $= 10$ dB			Source 2 SNR $= 5$ dB			Source 3 SNR $= 0$ dB		
r (m)	0.01	0.1	0.5	0.01	0.1	0.5	0.01	0.1	0.5
SRP (0.2 s)	100	90	76	99	89	63	89	71	35
Prop. (0.2 s)	100	100	100	100	99	99	90	89	87
SRP (0.7 s)	100	89	64	96	81	52	75	66	21
Prop. (0.7 s)	100	100	99	98	98	98	78	74	74

Table 4.2: Performance in Terms of Percentage of Correct Frames

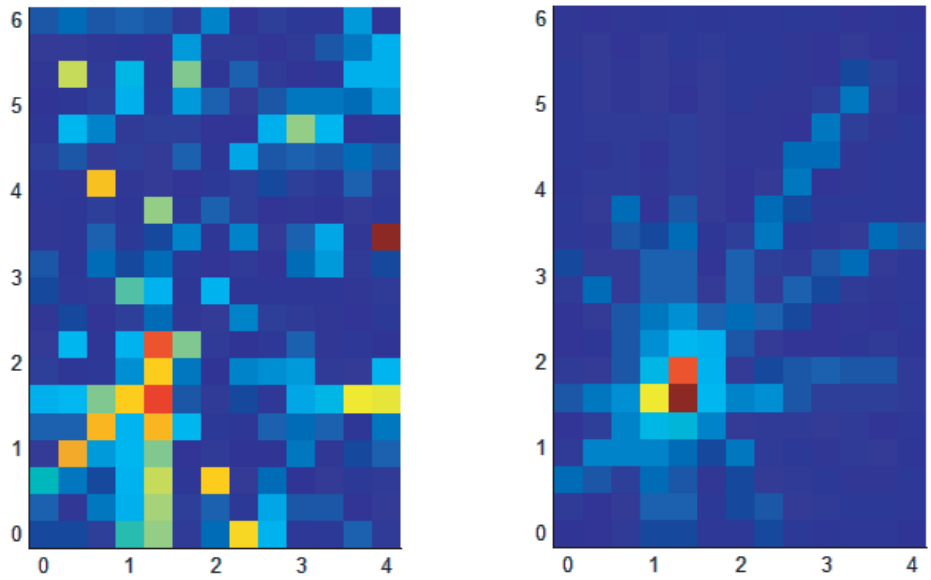


Figure 4.5: Source likelihood map. Coarse grid (a) traditional and (b) modified SRP-PHAT.

evaluations. Notice also that both methods perform almost the same in all situations when the finest grid is used.

Chapter 5

Speaker detection

A method for speaker detection based on the statistics of the resulting location estimates is provided in this section. The proposed speaker detection method is based on the probability density function of the location estimates by the improved SRP-PHAT algorithm explained in Chapter 3.

5.1 Speaker Detection

In the next subsections, we describe how active speakers are detected in our system, which requires a previous discrimination between speech and non-speech frames based on the distribution of location estimates. To this end, we model the probability density function of the obtained locations when there are active speakers and when silence and/or noise is present.

5.1.1 Distribution of Location Estimates

Our first step to speaker detection is to analyze the distribution of the location estimates $\hat{\mathbf{x}}_s$ when there is an active speaker talking inside the room from a static position. In this context, six microphones were placed on the walls of the videoconferencing room and a set of 12 recordings from different speaker positions were analyzed to obtain the resulting location estimates. Figure 5.1 shows an example of three two-dimensional histograms obtained from different speaker locations. It can be observed that, since the localization algorithm is very robust, the resulting distributions when speakers are active are significantly peaky. Also, notice that the shape of the distribution is very similar in all cases but centered in the actual speaker location. As a result, we model the distribution of estimates as a bivariate Laplacian as follows:

$$p(\hat{\mathbf{x}}_s|H_s(\mathbf{x}_s)) = \frac{1}{2\sigma_x\sigma_y}\exp^{-\sqrt{2}\left(\frac{|x-x_s|}{\sigma_x}+\frac{|y-y_s|}{\sigma_y}\right)}, \tag{5.1}$$

where $p(\hat{\mathbf{x}}_s|H_s(\mathbf{x}_s))$ is the conditional probability density function (pdf) of the location estimates under the hypothesis $H_s(\mathbf{x}_s)$ that there is an active speaker located at $\mathbf{x}_s = [x_s, y_s]$. Note that the variances σ_x^2 and σ_y^2 may depend on the specific microphone set-up and the selected processing parameters. This dependence will be addressed in future works.

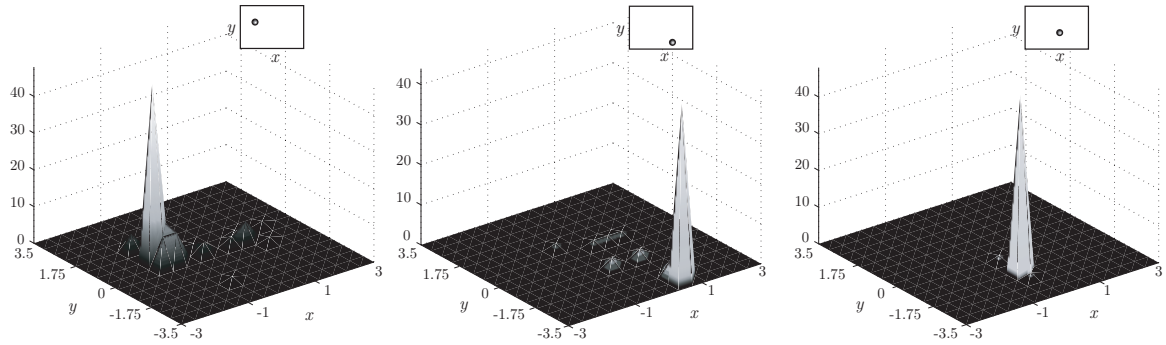


Figure 5.1: Distribution obtained for three different speaker locations.

On the other hand, a similar analysis was performed to study how the distribution changes when there are not active speakers, i.e. only noise frames are being processed. The resulting histogram can be observed in Figure 5.2, where it becomes apparent that the peakedness of this distribution is not as significant as the one obtained when there is an active source. Taking this into account, the distribution of non-speech frames is modeled as a bivariate Gaussian:

$$p(\hat{\mathbf{x}}_s|H_n) = \frac{1}{2\pi\sigma_{x_n}\sigma_{y_n}} \exp^{-\left(\frac{x^2}{2\sigma_{x_n}^2}+\frac{y^2}{2\sigma_{y_n}^2}\right)}, \tag{5.2}$$

where $p(\hat{\mathbf{x}}_s|H_n)$ is the conditional pdf of the location estimates under the hypothesis H_n that there are not active speakers, and the variances $\sigma_{x_n}^2$ and $\sigma_{y_n}^2$ are those obtained with noise-only frames.

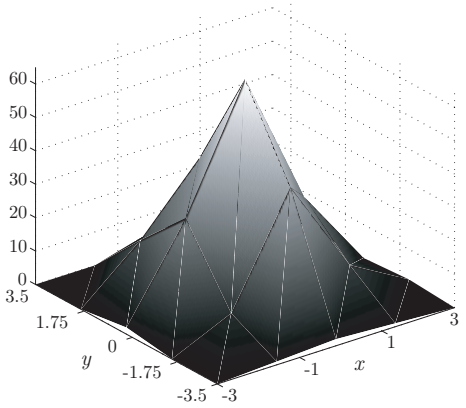


Figure 5.2: Distribution for non-speech frames.

5.1.2 Speech/Non-Speech Discrimination

In the above subsection, it has been shown that speech frames are characterized by a bivariate Laplacian probability density function. A similar analysis of location estimates when there are not active speakers results in a more Gaussian-like distribution, which is characterized by a shape less

peaky than a Laplacian distribution. This property is used in our system to discriminate between speech and non-speech frames by observing the peakedness of a set of accumulated estimates:

$$\mathbf{C} = \begin{bmatrix} \hat{x}_s(n) & \hat{y}_s(n) \\ \hat{x}_s(n-1) & \hat{y}_s(n-1) \\ \vdots & \vdots \\ \hat{x}_s(n-L-1) & \hat{y}_s(n-L-1) \end{bmatrix} = [\mathbf{c}_x\, \mathbf{c}_y], \tag{5.3}$$

where L is the number of the accumulated estimates in matrix $\mathbf{C}$. A peakedness criterion based on high-order statistics was evaluated. In probability theory and statistics, kurtosis is a measure of the "peakedness" of the probability distribution of a real-valued random variable.

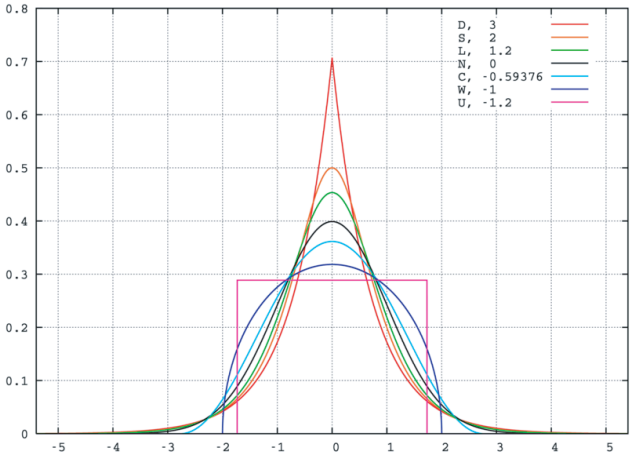


Figure 5.3: Excess Kurtosis for different density distributions.

Fig. 5.3 is an example where are compared several well-known distributions from different parametric families. All densities considered are unimodal and symmetric. Each has a mean and skewness of zero. Parameters were chosen to result in a variance of unity in each case. The seven densities are:

- **D**: Laplace distribution, red curve (two straight lines in the log-scale plot), excess kurtosis = 3
- **S**: hyperbolic secant distribution, orange curve, excess kurtosis = 2
- **L**: logistic distribution, green curve, excess kurtosis = 1.2
- **N**: normal distribution, black curve (inverted parabola in the log-scale plot), excess kurtosis = 0
- **C**: raised cosine distribution, cyan curve, excess kurtosis = -0.593762...
- **W**: Wigner semicircle distribution, blue curve, excess kurtosis = -1

- **U**: uniform distribution, magenta curve (shown for clarity as a rectangle in the image), excess kurtosis = -1.2.

Kurtosis is defined as a normalized form of the fourth central moment μ_4:

$$\text{Kurt}(\mathbf{c}_x) \equiv \frac{\mu_4}{\mu_2^2}, \tag{5.4}$$

where μ_i denotes the ith central moment (and in particular, μ_2 is the variance). The excess Kurtosis is defined by:

$$\gamma_2 \equiv \frac{\mu_4}{\mu_2^2} - 3, \tag{5.5}$$

Since the kurtosis of a normal distribution equals 3, we propose the following discrimination rules for active speech frames:

$$\text{Kurt}(\mathbf{c}_x) \begin{cases} \geq 3 & \text{speech} \\ < 3 & \text{non} - \text{speech} \end{cases}, \tag{5.6}$$

$$\text{Kurt}(\mathbf{c}_y) \begin{cases} \geq 3 & \text{speech} \\ < 3 & \text{non} - \text{speech} \end{cases}, \tag{5.7}$$

where a frame is selected as speech if any of the above conditions is fulfilled.

5.1.3 Camera Steering

To provide a suitable camera stability, a set of target positions were pre-defined coinciding with the actual seats in the videoconferencing room. The localization system will be responsible for communicating the camera which of the target positions is currently active. This process involves two main steps. First, it is necessary to discriminate between speech and non-speech frames as explained in Section 5.1.2. If a burst of speech frames is detected, then the estimated target position is forwarded to the camera when it does not match the current target seat. Since all the target positions are assumed to have the same prior probability, a maximum-likelihood criterion is followed:

$$\hat{\mathbf{x}}_t = \arg\max_{\mathbf{x}_t} p\left(\hat{\mathbf{x}}_s | H(\mathbf{x}_t)\right), \quad t = 1 \ldots N_t, \tag{5.8}$$

where $\mathbf{x}_t$ is one of the N_t pre-defined target positions. Given that the likelihoods have the same distribution centered at different locations, the estimated target position $\hat{\mathbf{x}}_t$ is the one which is closest to the estimated location $\hat{\mathbf{x}}_s$.

Chapter 6

Application to Videoconferencing

The SSL method explained in the chapter 3 has been applied in a videoconference system where, by accurately estimating the various users physical locations, it would be possible to steer a video camera toward the currently active speaker.

6.1 Set up for the videoconferece

To evaluate the performance of our proposed approach a set of recordings was carried out in a videoconferencing test room with dimensions 6.67 m x 5.76 m x 2.10 m. A set of 6 omnidirectional microphones were placed on the walls of the room.

To be precise, 4 of the microphones were situated at the 4 corners of the ceiling of the room and the other two microphones were placed at the same height but in the middle of the longest walls. Figure 6.1 shows the microphone set-up, the camera location and the different seats occupied by the participants. Black dots represent the 12 pre-defined target locations used to select the active speaker seat.

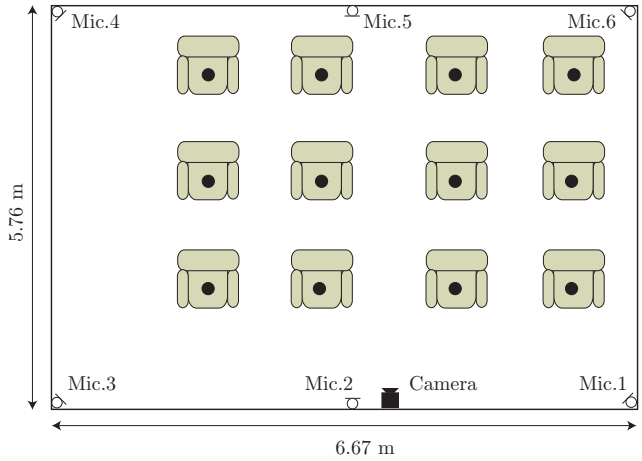


Figure 6.1: Room for the videoconference.

Grid res.	0.5 m				0.3 m			
L	5	10	15	20	5	10	15	20
% SP	52.5	60.4	70.0	74.0	68.9	70.7	83.1	85.4
% N-SP	75.9	64.8	70.9	72.7	81.4	70.9	81.5	82.3
% T	98.2				99.6			

Table 6.1: Performance in Terms of Percentage of Correct Frames

6.2 Description of the application

The experiment consisted in recording speakers talking from the different target positions (only one speaker at each time) with the corresponding space of silence between two talking interventions. The recordings were processed with the aim of evaluating the performance of our system in discriminating speech from non-speech frames and determining the active speaker so that the camera can point at the correct seat. With this aim, the original recordings were manually labeled as speech and non-speech fragments. The processing used a sampling rate of 44.1 kHz, with time windows of 2048 samples and 50% overlap. The location estimates were calculated using the modified SRP-PHAT functional, as explained in Chapter 3.

The discrimination between speech and non-speech frames was carried out by calculating the kurtosis of the last L estimated positions, as explained in Chapter 5.

6.3 Results

Modified SRP-PHAT approach joint the speech/non-speech discriminator have been used for a videoconference application, so different experiments have been carried out.

A new experiment was carried out in order to check the behavior of the speech/non-speech discriminator. To this aim, a set of recordings made in the test room (see Fig. 6.1) were used. These recordings were made from different pre-defined locations and they consist in active and non-active speakers, which is the same as people talking and noise environment.

Table 6.3 shows the percentage of correctly detected speech (% SP) and non-speech (% N-SP) frames with different number of accumulated positions $L = 5, 10, 15, 20$. Moreover, the processing was performed considering two different spatial grid sizes (0.3 m and 0.5 m). The percentage of speech frames with correct target positions (% T) is also shown in the table. It can be observed that, generally, the performance increases with a finer grid and with the number of accumulated estimates L. These results were expectable, since the involved statistics are better estimated with a higher number of location samples. Although it may seem that there are a significant number of speech frames that are not correctly discriminated, it should be noticed that this is not a problem for the correct driving of the camera, since most of them are isolated frames inside speech fragments that do not make the camera change its pointing target.

Chapter 7

Summary and Conclusions

Sound source localization and speech/non-speech detection techniques have been presented in this work to be used in a multiparticipant videoconferencing environment with a microphone array system for a steering-camera application.

Based on the well known SRP-PHAT SSL method, a modified version of that technique that uses a new functional has been developed. The proposed functional is based on the accumulation of GCC values in a range that covers the volume surrounding each point of the defined spatial grid. The GCC integration limits are determined by the gradient of the inter-microphone time delay function corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry. Our results showed that the proposed approach provides similar performance to the conventional SRP-PHAT algorithm in difficult environments with a reduction of five orders of magnitude in the required number of functional evaluations. This reduction has been shown to be sufficient for the development of real-time source localization applications.

In a videoconferencing environment where the sources are voices from different speakers, a speech/non-speech detection step is necessary to provide a robust steering camera system. For this reason the distribution of location estimates has been obtained using the proposed SRP-PHAT functional. Our analysis shows that location estimates follow different distributions when speakers are active or mute. This fact allows us to discriminate between speech and non-speech frames under a common localization framework. The results of experiments conducted in a real room suggest that, using a moderately high number of accumulated location estimates, it is possible to discriminate with significant accuracy between speech and non-speech frames, which is sufficient to correctly detect an active speaker and point the camera towards his/her predefined location.

To summarize, a modified SRP-PHAT algorithm for real-time SSL has been developed and evaluated in a practical scenario. The proposed method has been integrated into a speaker detection step to localize active sources in a videoconferencing room. In this context, a videocamera can be successfully driven by using the locations provided by our combined approach, showing the capabilities of the contributions described in this Master's thesis.

Chapter 8

Acknowledgments

This work was supported by the Ministry of Education and Science under the project TEC2009-14414-C03-01.

Bibliography

[1] AARABI, P. The fusion of distributed microphone arrays for sound localization. *EURASIP Journal on Applied Signal Processing 2003* (2003), 338–347.

[2] ALGHASSI, H. *Eye Array Sound Source Localization*. PhD thesis, University of British Columbia, 2008.

[3] CAMPBELL, D. R. Roomsim: a MATLAB simulation shoebox room acoustics, 2007. http://media.paisley.ac.uk/ campbell/Roomsim.

[4] CHA ZHANG, D., FLORENCIO, AND ZHENGYOU, Z. Why does PHAT work well in low noise, reverberant environments. ICASSP, pp. 2565–8.

[5] CHEN, J., BENESTY, J., AND HUANG, Y. Time delay estimation in room acoustic environments: an overview. *EURASIP Journal on Applied Signal Processing 2006* (2006), 1–19.

[6] COBOS, M., MARTI, A., AND LOPEZ, J. J. A modified srp-phat functional for robust real-time sound source localization with scalable spatial sampling. *IEEE Signal Processing Letters 18*, 1 (January 2011).

[7] DIBIASE, J. H. *A high accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. PhD thesis, Brown University, Providence, RI, May 2000.

[8] DO, H., AND SILVERMAN, H. F. A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC). In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)* (2007).

[9] DO, H., SILVERMAN, H. F., AND YU, Y. A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)* (2007).

[10] DONOHUE, K. D., HANNEMANN, J., AND DIETZ, H. G. Performance for phase transform for detecting sound sources in reverberant and noisy environments. *Signal Processing 87*, 7 (July 2007), 1677–1691.

[11] FRIEDLANDER, B., AND WEISS, A. J. Direction finding for wide-band signals using an interpolated array. *IEEE Transactions on Signal Processing* (April 1993), 41:1618:1634.

[12] JOHNSON, D. H., AND DUDGEON, D. E. *Array Signal Processing: Concepts and Techniques.* P T R Prentice Hall, 1993.

[13] KNAPP, C. H., AND CARTER, G. C. The generalized correlation method for estimation of time delay. *Transactions on Acoustics, Speech and Signal Processing ASSP-24* (1976), 320–327.

[14] MADHU, N., AND MARTIN, R. *Advances in Digital Speech Transmission*. Wiley, 2008, ch. Acoustic source localization with microphone arrays, pp. 135–166.

[15] MUNGAMURU, B. *Enhanced Sound Localization*. PhD thesis, University of Toronto, 2003.

[16] MUNGAMURU, B., AND AARABI, P. Enhanced sound localization. *IEEE Trans Syst, Man, Cybernet Part B: Cybernet 2004;34(3):152640*.

[17] PIRINEN, T. W. An experimental comparison of time delay weights for deirection of arrival estimation. *11th Int. Conference on Digital Audio Effects DAFx-08* (2008), 1–4.

[18] SILVERMAN, H. F., YU, Y., SACHAR, J. M., AND PATTERSON III, W. R. Performance of real-time source-location estimators for a large-aperture microphone array. *IEEE Transactions on Speech and Audio Processing 13* (2005), 593–606.

[19] STOICA, P., AND LI, J. Source localization from range-difference measurements. *IEEE Signal Processing Magazine* (November 2006), 63–69.

[20] SVAIZER, P., MATASSONI, M., AND OMOLOGO, M. Acoustic source location in a three-dimensional space cross-power spectrum phase. *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing ICASSP-97* (Munich, Germany, April 1997), 231–234.

[21] TELLAKULA, A. K. *Acoustic source localization using time delay estimation*. PhD thesis, Indiand Institute od Science, August 2007.

[22] TERVO, S., AND LOKKI, T. Interpolation methods for the srp-phat algorithm. *The 11th International Workshop on Acoustic Echo and Noise Control, Seattle, Washington, USA, IWAENC2008* (September 2008), 14–17.

[23] WANG, H., AND KAVEH, M. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wideband sources. *IEEE Transactions on Acoustics, Speech, and Signal Processing* (August 1985), ASSP–33:823:831.

[24] WATENABE, H., SUZUKI, M., NAGAI, N., AND MIKI, N. A method for maximum likelihood bearing estimation without nonlinear maximization. *Transactions of the Institute of Electronics, Information and Communication Engineers* (August 1989), J72A, 8:303:308.

Audio Engineering Society

Convention Paper

Presented at the 128th Convention
2010 May 22–25 London, UK

On the Effects of Room Reverberation in 3D DOA Estimation Using a Tetrahedral Microphone Array

Maximo Cobos[1], Jose J. Lopez[1] and Amparo Marti[1]

[1] *Institute of Telecommunications and Multimedia Applications (iTEAM), Universidad Politécnica de Valencia, Valencia, Camino de Vera s/n, 46022, Spain*

Correspondence should be addressed to Maximo Cobos (mcobos@iteam.upv.es)

ABSTRACT

This paper studies the accuracy in the estimation of the Direction-Of-Arrival (DOA) of multiple sound sources using a small microphone array. As other sparsity-based algorithms, the proposed method is able to work in underdetermined scenarios, where the number of sound sources exceeds the number of microphones. Moreover, the tetrahedral shape of the array allows to estimate DOAs in the 3-dimensional space easily, which is an advantage over other existing approaches. However, since the proposed processing is based on an anechoic signal model, the estimated DOA vectors are severely affected by room reflections. Experiments to analyze the resultant DOA distribution under different room conditions and source arrangements are discussed using both simulations and real recordings.

1. INTRODUCTION

Source localization is still one of the most challenging problems in acoustic signal processing. Estimating the *direction of arrival* (DOA) of multiple sound sources in a real scenario is a very difficult task. The estimation of DOAs of multiple sources has interesting applications in many speech processing systems, such as hands-free devices, teleconference systems or hearing aids. Algorithms for acoustic source localization are often classified into direct approaches and indirect approaches [1]. Indirect approaches estimate the *time difference of arrival* (TDOA) between various microphone pairs and then, based on the array geometry, estimate the source positions by optimization techniques. On the other hand, direct approaches compute a cost function over a set of

candidate locations and take the most likely source positions.

Cross-correlation-based methods, such as *Generalized Cross Correlation* (GCC) [2], are commonly applied in source localization. However, the GCC method becomes problematic when multiple sources are active simultaneously. Techniques based on the *Steered Response Power* (SRP) are also popular in acoustic source localization, but computationally demanding [3]. In the last years, localization methods based on the estimation of TDOAs in the time-frequency domain have been receiving increasing attention [4][5]. These algorithms provide considerable good accuracy with reduced computational complexity using the phase differences observed from two closely spaced sensors. However, their performance is considerably worse in non-anechoic environments, since room reflections affect the variance of DOA estimates. Therefore, source localization remains a very challenging task.

Recently, the authors studied the effect of room reflections in source localization using a small microphone array composed of three microphones [6], however, only the horizontal plane was considered. This paper discusses the accuracy achieved by a tetrahedral microphone array in 3-D source localization tasks. Following a sparsity-based approach, the microphone signals are transformed into the time-frequency domain. Then, phase-differences between microphone pairs are analyzed to provide an estimation of the DOA corresponding to the dominant source in each time-frequency bin. With the aim of discussing how the acoustic environment affects the distribution of DOA estimates in the 3-D space, a set of simulations considering different acoustic environments has been carried out. The results show that the statistics of the DOA vector norm provide a good description of the environment where the sound sources were recorded.

The paper is structured as follows. Section 2 presents the assumed signal model and the proposed processing used to estimate the location of several sound sources. Section 3 shows how the distribution of DOA estimates changes depending on the acoustic environment. Section 4 presents several experiments that analyze the statistical properties of DOA estimates using simulated rooms and real recordings. Finally, the conclusions of this work are summarized in Section 5.

2. SIGNAL MODEL AND DOA ESTIMATION

2.1. Signal Model

The signals recorded by a microphone array, with sensors denoted with indices $m = 1, 2, \ldots, M$ in an acoustic environment where N sound sources are present, can be modeled as a finite impulse response convolutive mixture, written as

$$x_m(t) = \sum_{n=1}^{N} \sum_{\ell=0}^{L_m-1} h_{mn}(\ell) s_n(t-\ell), \quad m = 1, \ldots, M \tag{1}$$

where $x_m(t)$ is the signal recorded at the m-th microphone at time sample t, $s_n(t)$ is the n-th source signal, $h_{mn}(t)$ is the impulse response of the acoustic path from source n to sensor m, and L_m is the maximum length of all impulse responses.

The above model can also be expressed in the *short-time Fourier transform* (STFT) domain as follows

$$X_m(k,r) = \sum_{n=1}^{N} H_{mn}(k) S(k,r), \tag{2}$$

where $X_m(k,r)$ denotes the STFT of the m-th microphone signal, being k and r the frequency index and time frame index, respectively. $S_n(k,r)$ denotes the STFT of the source signal $s_n(t)$ and $H_{mn}(k)$ is the frequency response from source n to sensor m.

2.1.1. Sparse Sources

In the time-frequency domain, source signals are usually assumed to be sparse. A sparse source has a peaky probability density function: the signal is close to zero at most time-frequency points, and has large values in rare occasions. This property has been widely applied in many works related to source signal localization [5][4] and separation [7][8] in underdetermined situations, i.e. when there are more sources than microphone signals.

If we assume that the sources rarely overlap at each time-frequency point, Equation (2) can be simplified as follows

$$X_m(k,r) \approx H_{ma}(k) S_a(k,r), \tag{3}$$

where $S_a(k,r)$ is the dominant source at time-frequency point (k,r). To simplify, we assume an anechoic model where the sources are sufficiently distant to consider plane wavefront incidence. Then, the frequency response is only a function of the time-delay τ_{mn} between each source and sensor

$$H_{mn}(k) = e^{\mathrm{j}2\pi f_k \tau_{mn}}, \tag{4}$$

being f_k the frequency corresponding to frequency index k.

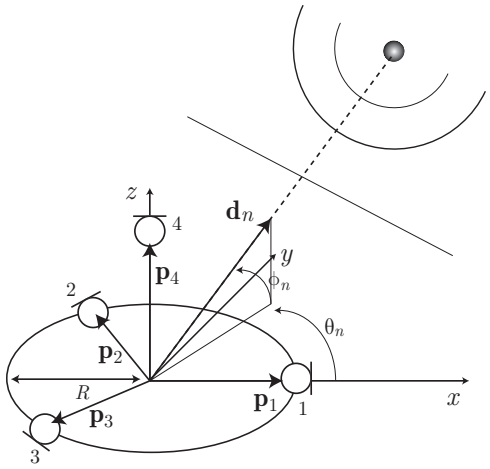


Fig. 1: Tetrahedral microphone array for 3-D DOA estimation.

2.2. Array Geometry and DOA Estimation

Now consider a tetrahedral microphone array ($M = 4$) with base radius R, as shown in Figure 1. The sensor location vectors in the 3-dimensional space with origin in the array base center, are given by:

$$\begin{aligned}
\mathbf{p}_1 &= [R,\ 0,\ 0]^T, \\
\mathbf{p}_2 &= \left[-\frac{R}{2},\ \frac{\sqrt{3}}{2}R,\ 0\right]^T, \\
\mathbf{p}_3 &= \left[-\frac{R}{2},\ -\frac{\sqrt{3}}{2}R,\ 0\right]^T, \\
\mathbf{p}_4 &= \left[0,\ 0,\ R\sqrt{2}\right]^T.
\end{aligned} \tag{5}$$

(6)

The DOA vector of the n-th source as a function of the azimuth θ_n and elevation ϕ_n angles is defined as

$$\mathbf{d}_n = [\cos\theta_n \cos\phi_n,\ \sin\theta_n \cos\phi_n,\ \sin\phi_n]^T. \tag{7}$$

The source to sensor time delay is given by $\tau_{mn} = \mathbf{p}_m^T \mathbf{d}_n / c$, being c the speed of sound. Therefore, the frequency response of Equation (4) can be written as

$$H_{mn}(k,r) \approx e^{\mathrm{j}\frac{2\pi f_k}{c}\mathbf{p}_m^T \mathbf{d}_n}. \tag{8}$$

Taking into account this last result and Equation 3, it becomes clear that the phase difference between the microphone pair formed by sensors i and j, is given by

$$\angle\left(\frac{X_j(k,r)}{X_i(k,r)}\right) \approx \frac{2\pi f_k}{c}(\mathbf{p}_j - \mathbf{p}_i)^T \mathbf{d}_n, \tag{9}$$

where $\angle$ denotes the phase of a complex number.

Using a reference microphone q, the phase difference information at point (k,r) of $M-1$ microphone pairs is stored in the vector

$$\mathbf{b}_q(k,r) = \left[\angle\left(\frac{X_1(k,r)}{X_q(k,r)}\right), \ldots, \angle\left(\frac{X_M(k,r)}{X_q(k,r)}\right)\right]^T, \tag{10}$$

forming the following system of equations:

$$\mathbf{b}_q(k,r) = \frac{2\pi f_k}{c}\mathbf{P}\mathbf{d}_n, \tag{11}$$

where

$$\mathbf{P} = [\mathbf{p}_{1q}, \ldots, \mathbf{p}_{Mq}]^T, \quad \mathbf{p}_{nq} = \mathbf{p}_n - \mathbf{p}_q. \tag{12}$$

Finally, the DOA at time-frequency bin (k,r) is obtained by taking the inverse of the $\mathbf{P}$ matrix

$$\hat{\mathbf{d}}_n(k,r) = \frac{c}{2\pi f_k}\mathbf{P}^{-1}\mathbf{b}_q(k,r). \tag{13}$$

The regular tetrahedral geometry used in this paper leads to the following simple equations for $\mathbf{d}_n(k,r) = [\hat{d}_1,\ \hat{d}_2,\ \hat{d}_3]^T$:

$$\hat{d}_1 = \cos\theta_n \cos\phi_n = \frac{c}{2\pi f_k}\frac{1}{\sqrt{3}}(b_2 + b_3), \tag{14}$$

$$\hat{d}_2 = \sin\theta_n \cos\phi_n = \frac{c}{2\pi f_k}(b_3 - b_2), \tag{15}$$

$$\hat{d}_3 = \sin\phi_n = \frac{c}{2\pi f_k}\left[\frac{1}{\sqrt{6}}(b_2 + b_3) - \sqrt{\frac{3}{2}}b_4\right] \tag{16}$$

where b_n is the n-th element of the vector $\mathbf{b}_1(k,r)$ (reference microphone $q = 1$). The azimuth angle

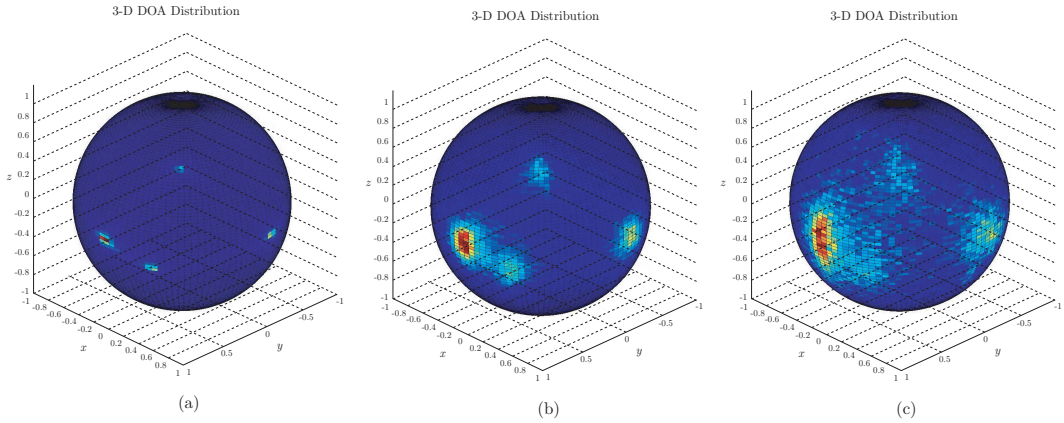


Fig. 2: Histograms showing the distribution of DOA estimates in the 3-D space calculated from a mixture of 4 speech sources. (a) Anechoic conditions. (b) $T_{60} = 150$ ms. (c) $T_{60} = 300$ ms.

is obtained using the four quadrant inverse tangent function:

$$\hat{\theta}_n(k,r) = \text{atan}^{360^\circ}(\hat{d}_1, \hat{d}_2). \tag{17}$$

The elevation angle is directly obtained as

$$\hat{\phi}_n(k,r) = \sin^{-1}(\hat{d}_3). \tag{18}$$

Note that for each time-frequency point (k,r), estimating the 3-D direction of arrival is relatively simple, just using the observed phase differences between 3 microphone pairs of the array. Another aspect to consider is spatial aliasing. The distance between microphones determines the angular aliasing frequency. Due to the 2π ambiguity in the calculation of the phase differences, the maximum ambiguity-free frequency in a microphone pair subarray would be given by $f_k = c/2d$, where d is the separation distance between the capsules. Beyond this frequency, there is no a one-to-one relationship between phase difference and spatial direction. However, small arrays with $d \approx 1.5$ cm provide an unambiguous bandwidth greater than 11 kHz, covering a perceptually important frequency range.

3. 3-D DOA DISTRIBUTIONS

The assumed signal model is close to reality when we are localizing in anechoic conditions. Obviously, the localization accuracy will be affected by room reflections when the localization task is performed in a reverberant environment. Moreover, room reflections also affect source sparseness [10] which is another basic assumption taken by the localization method.

In this section, we carry out some simulations considering a rectangular room and using different wall conditions in order to show how the distribution of DOA estimates is affected by room reflections.

3.1. Deviation of DOA estimates

With the objective of showing how the proposed array is capable of capturing the 3-D spatial information of sound, we show a simulated sound scene where 4 speech sources are simultaneously active in a room (10 s duration). The azimuth angles of the sources were $\theta_1 = 0^\circ$, $\theta_2 = 30^\circ$, $\theta_3 = 45^\circ$ and $\theta_4 = 100^\circ$. The elevation angles were $\phi_1 = 0^\circ$, $\phi_2 = 30^\circ$, $\phi_3 = -10^\circ$ and $\phi_4 = 45^\circ$. With the aim of showing graphically how the distribution of DOA estimates changes depending on the degree of reverberation, the sound scene was simulated using an increasing wall reflection factor [9], thus allowing more reflections inside the room. A more detailed description of the simulation set-up is given in Section 4. Figure 2 shows the 3-D histograms that represent the amount of estimates produced in a given direction. Note how in the anechoic case (a), the sources appear as localized peaky zones corresponding to their real DOAs. The diffuseness added by room reflections can be clearly seen in (b)-(c),

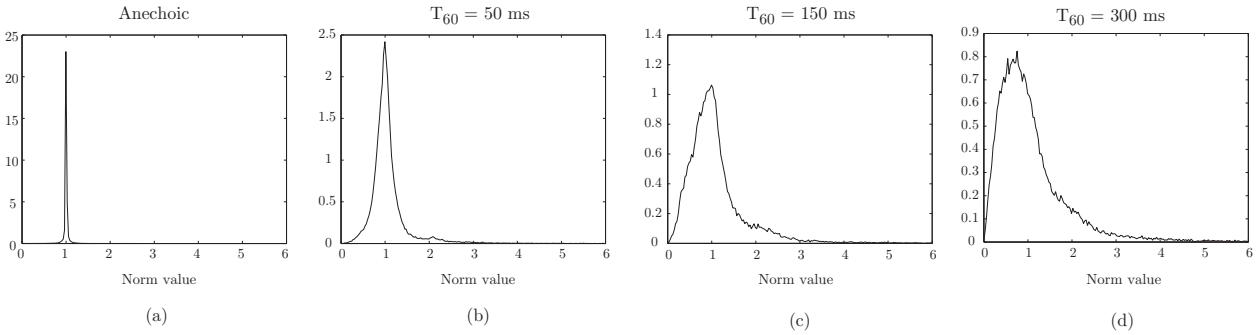


Fig. 3: Distribution of the DOA vector norm for different room conditions. (a) Anechoic room. (b) $T_{60} = 50$ ms. (c) $T_{60} = 150$ ms. (d) $T_{60} = 300$ ms.

where the estimates, although clustered around the real DOA directions, have been highly spread.

3.2. DOA vector norm distribution

It is important to note that perfectly estimated directions will have unit norm, i.e. $||\hat{\mathbf{d}}_n(k,r)|| = 1$. Therefore, perfect estimations fulfilling the used anechoic model will lie on the unit sphere. In contrast, the norm of the estimated DOA vector in points with high spectral overlap between the sources or corrupted by reverberation will be further away from the unity. Figure 3 shows four examples of norm distributions obtained from different simulated rooms with a single active source located at $\theta = 0°$, $\phi = 0°$. Note that in the anechoic case, the resultant norm distribution has a very large peak in the unity, whereas in the case of reverberant rooms, the distribution is substantially spread and asymmetric.

In the next section, we will study in detail the effect that the source-to-array distance and the number of sources have in the DOA vector norm distribution for different room conditions.

4. EXPERIMENTS

As shown in the last section, the presence of room reflections has a considerable effect on the estimated DOA vectors, since the anechoic signal model becomes corrupted with reverberation. Thus, the direct path contribution is very important to obtain correct DOA estimates. Moreover, the sparseness assumption also becomes affected by reverberation and by the number of sources. In this section, we conduct a set of experiments focused on the statistical analysis of the resultant DOA norm distribution under different situations.

4.1. Simulations

Several sound scenes have been simulated to discuss some important aspects previously commented: reverberation time, number of sources and direct-path contribution. In the simulations, a set of sound sources were positioned inside a shoe-box-shaped room (4 m × 3.6 m × 2.6 m) and all the source-to-sensor impulse responses were acquired by means of the mirror image method [9]. The wall reflection factor of the walls was changed to get different reverberation times (anechoic, $T_{60} = 50$ ms, $T_{60} = 150$ ms and $T_{60} = 300$ ms). Different number of sound sources (speech) were considered:

- 1 source at ($\theta = 0°$, $\phi = 0°$).
- 2 sources at
  1. ($\theta_1 = 45°$, $\phi_1 = 0°$),
  2. ($\theta_2 = -45°$, $\phi_2 = 0°$).
- 4 sources at
  1. ($\theta_1 = 0°$, $\phi_1 = 0°$),
  2. ($\theta_2 = 90°$, $\phi_2 = -30°$),
  3. ($\theta_3 = 180°$, $\phi_3 = 30°$),
  4. ($\theta_4 = -90°$, $\phi_4 = 60°$).

Moreover, different distances from the sources to the array were taken into account to modify the

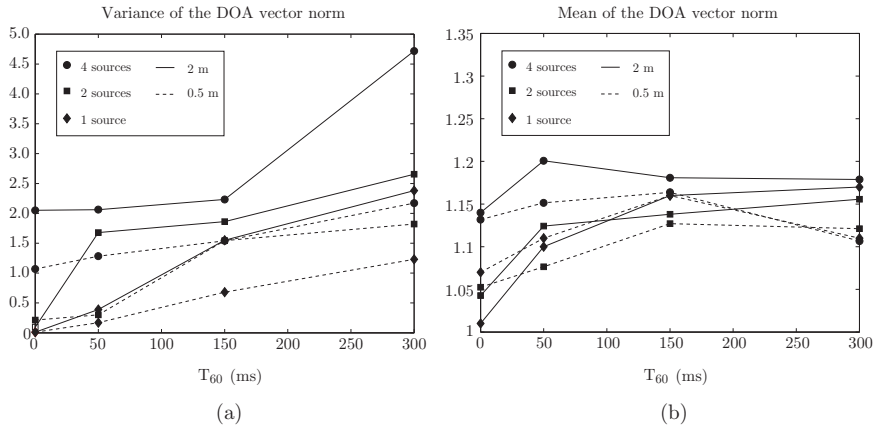


Fig. 4: Variance and mean of the DOA vector norm for different simulated environments. (a) Variance. (b) Mean.

direct path contribution in the same room. Thus, the simulations were carried out considering sources 2 m away from the array and 0.5 m away from it. The signals were sampled at 16 kHz, and the STFT processing was done by using Hann-windowed time frames of 1024 samples and 50% overlap.

Figure 4(a) shows the variance of the norm of the DOA vector estimates $||\hat{\mathbf{d}}_n(k,r)||$ for different rooms, source-to array distance and number of sources. As expected, the best case is that of a single source in an anechoic room and close to the tetrahedral array. When reverberation appears, the variance found in the estimates is greater, being more important the change when the number of sources and/or their distance to the array is increased.

Figure 4(b) shows the mean of the norm of the DOA vector estimates for the same cases. Similarly to the variance, the mean tends to be closer to unity in situations with less reverberation, less sources, and less source-to-sensor distance. However, the changes produced in the mean are not as large as in the case of the variance.

4.2. Real Room

Some preliminary experiments in a real room using a tetrahedral microphone prototype were conducted. The signal acquiring system consisted of a digital audio interface with four microphone inputs (M-Audio Fast Track Ultra USB 2.0). To construct the microphone array prototype, four instrumentation quality microphones from Earthworks model M-30 were used. These microphones have an almost perfectly planar response ($\pm$0.5 dB) in the audio band, and a very accurate phase match until high frequencies.

In this experiment, two subjects talking were recorded (10 s duration) in our recording studio, which has a reverberation time of approximately 0.2 s. These subjects were positioned at a distance of 0.5 m from the array, following the same arrangement as in the two-source simulations. The variance and mean from the norm of the DOA vectors obtained after processing the four microphone signals were 1.53 and 1.1, respectively. This result is in agreement with the values obtained in the previous simulations, confirming the relationship between these simple statistics and the acoustic environment.

5. CONCLUSIONS

Localization methods based on source sparseness in the time-frequency domain have been receiving increasing attention in the last years. Following the principles of sparsity-based localization methods, we presented a small tetrahedral microphone array that is capable of localizing several sound sources in the 3-D space. However, the assumed signal model is only close to reality when localization is performed in anechoic conditions. Thus, DOA estimates are severely affected by room reflections when working inside a reverberant environment. With the aim of characterizing the accuracy of DOA estimates, the

distribution of the estimated DOA vector norm was studied in different acoustic environments. Specifically, several experiments were conducted to assess the effect that the source-to-array distance and the number of sources have in the DOA vector norm distribution under different room conditions. The results showed that, for a certain reverberation time, the variance of the DOA vector norm is substantially increased when the number of sources and their distance to the array becomes larger. Our future work will be centered on using these preliminary results to develop a model that characterizes the acoustic environment by using the information extracted from a source localization system.

6. ACKNOWLEDGMENTS

The Spanish Ministry of Science and Innovation supported this work under the project TEC2009-14414-C03-01.

7. REFERENCES

[1] N. Madhu and R. Martin, "Advances in Digital Speech Transmission," Wiley, 2008, pp. 135166.

[2] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-24, pp. 320327, 1976.

[3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Microphone Arrays: Signal Processing Techniques and Applications," Springer- Verlag, 2001, ch. Robust Localization in Reverberant Rooms, pp. 157 180.

[4] S. Rickard and F. Dietrich,"DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET," in *Proceedings of the 10th IEEE Workshop on Statistical Signal and Array Processing (SSAP2000)*, Pocono Manor, PA, August 2000, pp. 311314.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors," *Journal of Signal Processing Systems*, 2009.

[6] M. Cobos, J. J. Lopez and S. Spors, "Analysis of room reverberation effects in source localization using small microphone arrays," presented at the *International Symposium on Communications, Control and Signal Processing (ISCCSP 2010)*, Limassol, Cyprus, March 2010.

[7] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," in *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 18301847, July 2004.

[8] P. Bofill and M. Zibulevski, "Underdetermined blind source separation using sparse representations," in *Signal Processing*, vol. 81, pp. 2353-2362, 2001.

[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," in *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943950, 1979.

[10] S. Rickard and O. Yilmaz, "On the W-disjoint orthogonality of speech," presented at *the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2002)*, Orlando, Florida, USA, 2002, pp.529-532.

A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling

Maximo Cobos, *Member, IEEE*, Amparo Marti, *Student Member, IEEE*, and Jose J. Lopez, *Member, IEEE*

***Abstract*—The Steered Response Power – Phase Transform (SRP-PHAT) algorithm has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. However, its practical implementation is usually based on a costly fine grid-search procedure, making the computational cost of the method a real issue. In this letter, we introduce an effective strategy that extends the conventional SRP-PHAT functional with the aim of considering the volume surrounding the discrete locations of the spatial grid. As a result, the modified functional performs a full exploration of the sampled space rather than computing the SRP at discrete spatial positions, increasing its robustness and allowing for a coarser spatial grid. To this end, the Generalized Cross-Correlation (GCC) function corresponding to each microphone pair must be properly accumulated according to the defined microphone setup. Experiments carried out under different acoustic conditions confirm the validity of the proposed approach.**

***Index Terms*—Microphone array, sound source localization, SRP-PHAT.**

I. Introduction

SOUND source localization under high noise and reverberation still remains a very challenging task. To this end, microphone arrays are commonly employed in many sound processing applications such as videoconferencing, hands-free speech acquisition, digital hearing aids, video-gaming, autonomous robots and remote surveillance. Algorithms for sound source localization can be broadly divided into indirect and direct approaches [1]. Indirect approaches usually follow a two-step procedure: they first estimate the *Time Difference Of Arrival* (TDOA) [2] between microphone pairs and, afterwards, they estimate the source position based on the geometry of the array and the estimated delays. On the other hand, direct approaches perform TDOA estimation and source localization in one single step by scanning a set of candidate source locations and selecting the most likely position as an estimate of the source location. In addition, information theoretic approaches have also shown to be significantly powerful in source localization tasks [3].

The *Steered Response Power – Phase Transform* (SRP-PHAT) algorithm is a direct approach that has been shown to be very robust under difficult acoustic conditions [4]–[6]. The algorithm is commonly interpreted as a beamforming-based approach that searches for the candidate source position that maximizes the output of a steered delay-and-sum beamformer. However, despite its robustness, computational cost is a real issue because the SRP space to be searched has many local extrema [7]. Very interesting modifications and optimizations have already been proposed to deal with this problem, such as those based on Stochastic Region Contraction (SRC) [8] and coarse-to-fine region contraction [9], achieving a reduction in computational cost of more than three orders of magnitude.

In this letter, we propose a different strategy where, instead of evaluating the SRP functional at discrete positions of a spatial grid, it is accumulated over the *Generalized Cross Correlation* (GCC) lag space corresponding to the volume surrounding each point of the grid. The GCC accumulation limits are determined by the gradient of the inter-microphone time delay function corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry. The benefits of following this approach are twofold. On the one hand, it incorporates additional spatial knowledge at each point for making a better final decision. On the other hand, the proposed modification achieves the same performance as SRP-PHAT with fewer functional evaluations, relaxing the computational demand required for a practical application.

II. The SRP-PHAT Algorithm

Consider the output from microphone l, $m_l(t)$, in an M microphone system. Then, the SRP at the spatial point $\mathbf{x} = [x, y, z]$ for a time frame n of length T is defined as

$$P_n(\mathbf{x}) \equiv \int_{nT}^{(n+1)T} \left| \sum_{l=1}^{M} w_l m_l \left(t - \tau(\mathbf{x}, l)\right) \right|^2 dt, \qquad (1)$$

where w_l is a weight and $\tau(\mathbf{x}, l)$ is the direct time of travel from location $\mathbf{x}$ to microphone l. DiBiase [7] showed that the SRP can be computed by summing the GCCs for all possible pairs of the set of microphones. The GCC for a microphone pair (k, l) is computed as

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \Phi_{kl}(\omega) M_k(\omega) M_l^*(\omega) e^{j\omega\tau} d\omega \qquad (2)$$

Manuscript received September 06, 2010; revised October 22, 2010; accepted October 27, 2010. This work was suported by the The Spanish Ministry of Science and Innovation supported this work under the project TEC2009-14414-C03-01. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Constantine L. Kotropoulos.

The authors are with the Institute of Telecommunications and Multimedia Applications, Universidad Politécnica de Valencia, 46022 Valencia, Spain (e-mail: mcobos@iteam.upv.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2010.2091502

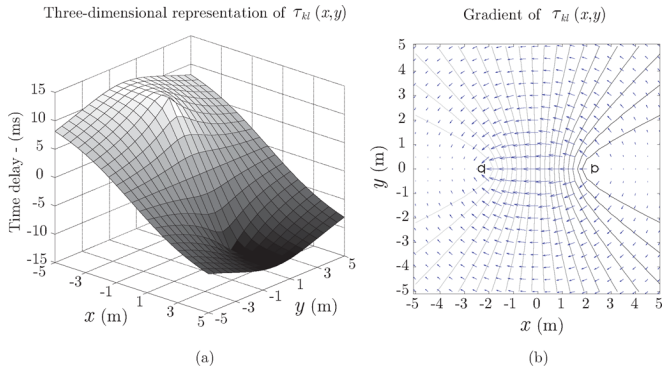


Fig. 1. Example of IMTDF. (a) Representation for the plane $z = 0$ with microphones located at $[-2, 0, 0]$ and $[2, 0, 0]$. (b) Gradient.

where τ is the time lag, * denotes complex conjugation, $M_l(\omega)$ is the Fourier transform of the microphone signal $m_l(t)$, and $\Phi_{kl}(\omega)$ is a combined weighting function in the frequency domain. The phase transform (PHAT) [10] has been demonstrated to be a very effective GCC weighting for time delay estimation in reverberant environments:

$$\Phi_{kl}(\omega) \equiv \frac{1}{|M_k(\omega)M_l^*(\omega)|}. \tag{3}$$

Taking into account the symmetries involved in the computation of (1) and removing some fixed energy terms [7], the part of $P_n(\mathbf{x})$ that changes with $\mathbf{x}$ is isolated as

$$P'_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} R_{m_k m_l}\left(\tau_{kl}(\mathbf{x})\right) \tag{4}$$

where $\tau_{kl}(\mathbf{x})$ is the *inter-microphone time-delay function* (IMTDF). This function is very important, since it represents the theoretical direct path delay for the microphone pair (k, l) resulting from a point source located at $\mathbf{x}$. The IMTDF is mathematically expressed as

$$\tau_{kl}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{x}_k\| - \|\mathbf{x} - \mathbf{x}_l\|}{c} \tag{5}$$

where c is the speed of sound, and $\mathbf{x}_k$ and $\mathbf{x}_l$ are the microphone locations.

The SRP-PHAT algorithm consists in evaluating the functional $P'_n(\mathbf{x})$ on a fine grid G with the aim of finding the point-source location $\mathbf{x}_s$ that provides the maximum value:

$$\mathbf{x}_s = \arg\max_{\mathbf{x}\in G} P'_n(\mathbf{x}). \tag{6}$$

III. The Inter-Microphone Time Delay Function

As commented in the previous section, the IMTDF plays a very important role in the source localization task. This function can be interpreted as the spatial distribution of possible TDOAs resulting from a given microphone pair geometry.

The function $\tau_{kl}(\mathbf{x})$ is continuous in $\mathbf{x}$ and changes rapidly at points close to the line connecting both microphone locations. Therefore, a pair of microphones used as a time-delay sensor is maximally sensible to changes produced over this line [11]. An example function is depicted in Fig. 1(a) for the plane $z = 0$, with $\mathbf{x}_k = [-2, 0, 0]$ and $\mathbf{x}_l = [2, 0, 0]$. The gradient of the function is shown in Fig. 1(b).

It is useful here to remark that the equation $|\tau_{kl}(\mathbf{x})| = C$, with C being a positive real constant, defines a hyperboloid in space with foci on the microphone locations $\mathbf{x}_k$ and $\mathbf{x}_l$. Moreover, the set of continuous confocal half-hyperboloids $\tau_{kl}(\mathbf{x}) = C$ with $C \in [-C_{\max}, C_{\max}]$, being $C_{\max} = (1/c)\|\mathbf{x}_k - \mathbf{x}_l\|$, spans the whole 3-D space.

Theorem: Given a volume V in space, the IMTDF for points inside V, $\tau_{kl}(\mathbf{x} \in V)$, takes only values in the continuous range $[\min(\tau_{kl}(\mathbf{x} \in \partial V)), \max(\tau_{kl}(\mathbf{x} \in \partial V))]$, where ∂V is the boundary surface that encloses V.

Proof: Let us assume that a point inside V, $\mathbf{x}_0 \in V$, takes the maximum value in the volume, i.e., $\tau_{kl}(\mathbf{x}_0) = \max(\tau_{kl}(\mathbf{x} \in V)) = C_{\max_V}$. Since there is a half-hyperboloid that goes through each point of the space, all the points besides $\mathbf{x}_0$ satisfying $\tau_{kl}(\mathbf{x}) = C_{\max_V}$ will also take the maximum value. Therefore, all the points on the surface resulting from the intersection of the volume and the half-hyperboloid will take this maximum value, including those pertaining to the boundary surface ∂V. The existence of the minimum in ∂V is similarly deduced.

The above property is very useful to understand the advantages of the approach presented in this letter. Note that the SRP-PHAT algorithm is based on accumulating the values of the different GCCs at those time lags coinciding with the theoretical inter-microphone time delays, which are only computed at discrete points of a spatial grid. However, as described before, it is possible to analyze a complete spatial volume by scanning the time-delays contained in a range defined by the maximum and minimum values on its boundary surface. In the next section, we describe how this knowledge can be included in the localization algorithm to increase its robustness.

IV. Proposed Approach

Let us begin the description of the proposed approach by analyzing a simple case where we want to estimate the location $\mathbf{x}_s$ of a sound source inside an anechoic space. In this simple case, the GCCs corresponding to each microphone pair are delta functions centered at the corresponding inter-microphone time-delays: $R_{m_k m_l}(\tau) = \delta(\tau - \tau_{kl}(\mathbf{x}_s))$. For example and without loss of generality, let us assume a setup with $M = 3$ microphones, as depicted in Fig. 2(a). Then, the source position would be that of the intersection of the three half-hyperboloids $\tau_{kl}(\mathbf{x}) = \tau_{kl}(\mathbf{x}_s)$, with $(k, l) \in \{(1, 2), (1, 3), (2, 3)\}$. Consider now that, to localize the source, a spatial grid with resolution $r = 1$ m is used as shown in Fig. 2(a). Unfortunately, the intersection does not match any of the sampled positions, leading to an error in the localization task. Obviously, this problem would have been easier to solve with a two step localization approach, but the above example shows the limitations imposed by the selected spatial sampling in SRP-PHAT, even in optimal acoustic conditions. This is not the case of the approach followed to localize the source in Fig. 2(b) where, using the same spatial grid, the GCCs have been integrated for each sampled position in a range that covers their volume of influence. A darker gray color indicates a greater accumulated value and, therefore, the darkest area is being correctly identified as the one containing the true sound source location. This new modified functional is expressed as follows

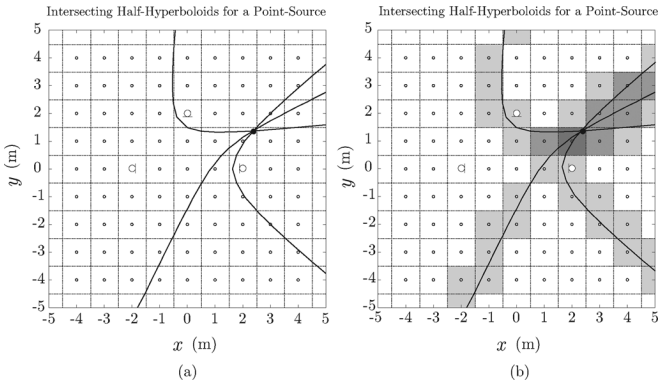


Fig. 2. Intersecting half-hyperboloids and localization approaches. (a) Conventional SRP-PHAT. (b) Proposed.

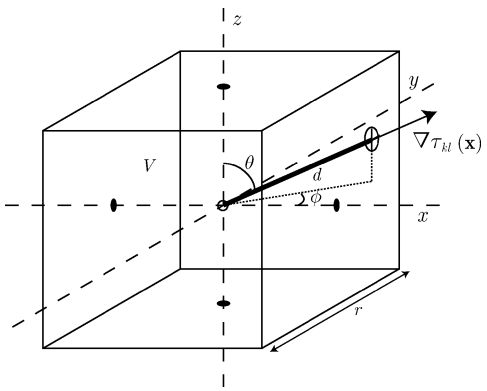


Fig. 3. Volume of influence of a point in a rectangular grid.

$$P''_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \sum_{\tau=L_{kl1}(\mathbf{x})}^{L_{kl2}(\mathbf{x})} R_{m_k m_l}(\tau). \tag{7}$$

The problem is to determine correctly the limits $L_{kl1}(\mathbf{x})$ and $L_{kl2}(\mathbf{x})$, which depend on the specific IMTDF resulting from each microphone pair. The computation of these limits is explained in the next subsection.

A. Computation of Accumulation Limits

As explained in Section III, the IMTDF inside a volume can only take values in the range defined by its boundary surface. Therefore, for each point of the grid, the problem of finding the GCC accumulation limits of its volume of influence can be simplified to finding the maximum and minimum values on the boundary. To this end, it becomes useful to study the direction of the greatest rate of increase at each grid point, which is given by the gradient

$$\nabla\tau_{kl}(\mathbf{x}) = [\nabla_x\tau_{kl}(\mathbf{x}), \nabla_y\tau_{kl}(\mathbf{x}), \nabla_z\tau_{kl}(\mathbf{x})] \tag{8}$$

where each component of the gradient vector can be calculated with

$$\nabla_\gamma\tau_{kl}(\mathbf{x}) = \frac{\partial\tau_{kl}(\mathbf{x})}{\partial\gamma} = \frac{1}{c}\left(\frac{\gamma-\gamma_k}{\|\mathbf{x}-\mathbf{x}_k\|} - \frac{\gamma-\gamma_l}{\|\mathbf{x}-\mathbf{x}_l\|}\right) \tag{9}$$

where γ denotes either x, y or z. The accumulation limits for a symmetric volume surrounding a point of the grid can be calculated by taking the product of the magnitude of the gradient and the distance d that exists from the point to the boundary following the gradient's direction:

$$L_{kl1}(\mathbf{x}) = \tau_{kl}(\mathbf{x}) - \|\nabla\tau_{kl}(\mathbf{x})\| \cdot d, \tag{10}$$

$$L_{kl2}(\mathbf{x}) = \tau_{kl}(\mathbf{x}) + \|\nabla\tau_{kl}(\mathbf{x})\| \cdot d \tag{11}$$

Fig. 3 depicts the geometry for a rectangular grid with spatial resolution r. For this cubic geometry, the distance d can be expressed as

$$d = \frac{r}{2}\min\left(\frac{1}{|\sin\theta\cos\phi|}, \frac{1}{|\sin\theta\sin\phi|}, \frac{1}{|\cos\theta|}\right) \tag{12}$$

where

$$\theta = \cos^{-1}\left(\frac{\nabla_z\tau_{kl}(\mathbf{x})}{\|\nabla\tau_{kl}(\mathbf{x})\|}\right), \tag{13}$$

$$\phi = \mathrm{atan}_2(\nabla_y\tau_{kl}(\mathbf{x}), \nabla_x\tau_{kl}(\mathbf{x})) \tag{14}$$

being $\mathrm{atan}_2(y,x)$ the quadrant-resolving arctangent function.

B. Computational Cost

Let L be the DFT length of a frame and $Q = M(M-1)/2$ the number of microphone pairs. The computational cost of SRP-PHAT is given by [5]:

$$\mathrm{SRP-PHAT}_{cost} \approx [6.125Q^2 + 3.75Q]L\log_2 L + 15LQ(1.5Q-1) + (45Q^2 - 30Q)\nu' \tag{15}$$

where ν' is the average number of functional evaluations required to find the maximum of the SRP space. Since the cost added by the modified functional is negligible and the frequency-domain processing of our approach remains the same as the conventional SRP-PHAT algorithm, the above formula is valid for both approaches. Moreover, since the accumulation limits can be precomputed before running the localization algorithm, the associated processing does not involve additional computation effort. However, as it will be shown in the next subsection, the advantage of the proposed method relies on the reduced number of required functional evaluations ν' for detecting the true source location, which results in an improved computational efficiency.

V. Experiments

Different experiments with real and synthetic recordings were conducted to compare the performances of the conventional SRP-PHAT algorithm, the SRC algorithm and our proposed method. First, the *Roomsim* Matlab package [12] was used to simulate an array of six microphones placed on the walls of a shoe-box-shaped room with dimensions 4 m × 6 m × 2 m (Fig. 4(a)). The simulations were repeated with two different reverberation times ($\mathrm{T}_{60} = 0.2$ s and $\mathrm{T}_{60} = 0.7$ s), considering 30 random source locations and different signal-to-noise ratio (SNR) conditions. The resultant recordings were processed with three different spatial grid resolutions in the case of SRP-PHAT and the proposed method ($r_1 = 0.01$ m, $r_2 = 0.1$ m and $r_3 = 0.5$ m). Note that the number of functional evaluations

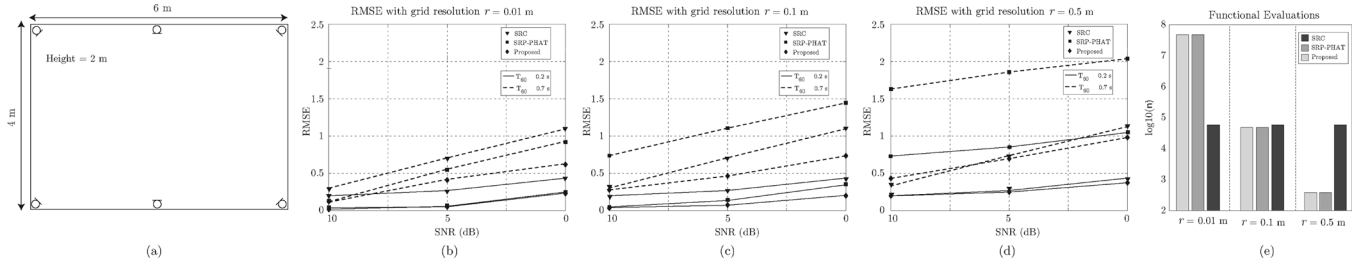


Fig. 4. Results with simulations. (a) Setup. (b) $r = 0.01\,\mathrm{m}$. (c) $r = 0.1\,\mathrm{m}$. (d) $r = 0.5\,\mathrm{m}$. (e) Functional evaluations.

TABLE I
RMSE FOR THE REAL-DATA EXPERIMENT

r	0.01	0.1	0.5
ν'	$802 \cdot 10^5$	$802 \cdot 10^2$	641
SRP-PHAT	RMSE = 0.29	RMSE = 0.74	RMSE = 1.82
Proposed	RMSE = 0.21	RMSE = 0.29	RMSE = 0.31
SRC	RMSE = 0.34 ($\nu' = 58307$)		

ν' depends on the selected value of r, having $\nu'_1 = 480 \times 10^5$, $\nu'_2 = 480 \times 10^2$ and $\nu'_3 = 384$. The implementation of SRC was the one made available by Brown University's LEMS at http://www.lems.brown.edu/array/download.html, using 3000 initial random points. The processing was carried out using a sampling rate of 44.1 kHz, with time windows of 4096 samples of length and 50% overlap. The simulated sources were male and female speech signals of length 5 s with no pauses. The averaged results in terms of *Root Mean Squared Error* (RMSE) are shown in Fig. 4(b)–(d). Since SRC does not depend on the grid size, the SRC curves are the same in all these graphs. As expected, all the tested systems perform considerably better in the case of low reverberation and high SNR. For the finest grid, it can be clearly observed that the performance of SRP-PHAT and the proposed method is almost the same. However, for coarser grids, our proposed method is only slightly degraded, while the performance of SRP-PHAT becomes substantially worse, specially for low SNRs and high reverberation. SRC has similar performance to SRP-PHAT with $r = 0.01$ m. Therefore, our proposed approach performs robustly with higher grid sizes, which results in a great computational saving in terms of functional evaluations, as depicted in Fig. 4(e).

On the other hand, a real setup quite similar to the simulated one was considered to study the performance of the method in a real scenario. Six omnidirectional microphones were placed at the four corners and at the middle of the longest walls of a video-conferencing room with dimensions 5.7 m $\times$ 6.7 m $\times$ 2.1 m and 12 seats. The measured reverberation time was $T_{60} = 0.28$ s. The processing was the same as with the synthetic recordings, using continuous speech fragments obtained from the 12 seat locations. The results are shown in Table I and confirm that our proposed method performs robustly using a very coarse grid. Although similar accuracy to SRC is obtained, the number of functional evaluations is significantly reduced.

VI. CONCLUSION

This letter presented a robust approach to sound source localization based on a modified version of the well-known SRP-PHAT algorithm. The proposed functional is based on the accumulation of GCC values in a range that covers the volume surrounding each point of the defined spatial grid. The GCC accumulation limits are determined by the gradient of the inter-microphone time delay function corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry. Our results showed that the proposed approach provides similar performance to the conventional SRP-PHAT algorithm in difficult environments with a reduction of five orders of magnitude in the required number of functional evaluations, with further computational saving than SRC. This reduction has been shown to be sufficient for the development of real-time source localization applications.

REFERENCES

[1] N. Madhu and R. Martin, "Acoustic source localization with microphone arrays," in *Advances in Digital Speech Transmission*. Hoboken, NJ: Wiley, 2008, pp. 135–166.
[2] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–19, 2006.
[3] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos, "Estimation of direction of arrival using information theory," *IEEE Signal Process.*, vol. 12, no. 8, pp. 561–564, 2005.
[4] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001, pp. 157–180.
[5] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson, III, "Performance of real-time source-location estimators for a large-aperture microphone array," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 593–606, 2005.
[6] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 4, pp. 338–347, 2003.
[7] J. H. DiBiase, "A High Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments using Microphone Arrays," Ph.D. dissertation, Brown Univ., Providence, RI, 2000.
[8] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2007)*, Honolulu, HI, Apr. 2007.
[9] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)*, New Paltz, NY, Oct. 2007.
[10] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *Trans. Acoust. Speech, Signal Process.*, vol. ASSP-24, pp. 320–327, 1976.
[11] E. A. P. Habets and P. C. W. Sommen, "Optimal microphone placement for source localization using time delay estimation," in *Proc. 13th Annu. Workshop on Circuits, Systems and Signal Processing (ProRISC 2002)*, Veldhoven, The Netherlands, 2002.
[12] D. R. Campbell, Roomsim: A MATLAB Simulation Shoebox Room Acoustics 2007 [Online]. Available: http://media.paisley.ac.uk/~campbell/Roomsim

REAL TIME SPEAKER LOCALIZATION AND DETECTION SYSTEM FOR CAMERA STEERING IN MULTIPARTICIPANT VIDEOCONFERENCING ENVIRONMENTS

Amparo Marti, Maximo Cobos, Jose J. Lopez

Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM),
Universidad Politécnica de Valencia

ABSTRACT

A real time speaker localization and detection system for videoconferencing environments is presented. In this system, a recently proposed modified *Steered Response Power - Phase Transform* (SRP-PHAT) algorithm has been used as the core processing scheme. The new SRP-PHAT functional has been shown to provide robust localization performance in indoor environments without the need for having a very fine spatial grid, thus reducing the computational cost required in a practical implementation. Moreover, it has been demonstrated that the statistical distribution of location estimates when a speaker is active can be successfully used to discriminate between speech and non-speech frames by using a criterion of peakedness. As a result, talking participants can be detected and located with significant accuracy following a common processing framework.

Index Terms— SRP-PHAT, source localization, speaker detection, microphone arrays

1. INTRODUCTION

Many applications, ranging from teleconferencing systems to artificial perception, hands-free speech acquisition, digital hearing aids, video-gaming, autonomous robots and remote surveillance require the localization of one or more acoustic sources. Since the boost of new generation videoconferencing environments, there has been growing interest in the development of automatic camera-steering systems using microphone arrays [1],[2]. In this work, we present a microphone array system for camera-steering to be used in a multiparticipant videoconferencing environment based on the well-known SRP-PHAT algorithm [3]. The SRP-PHAT method has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. It is commonly interpreted as a beamforming-based approach that searches for the candidate source position that maximizes the output of a steered delay-and-sum beamformer. However, the computational requirements of the method are large, making its real-time implementation considerably difficult. Since the SRP-PHAT method was proposed, there have been several attempts to reduce the computational cost of the method, such as those presented in [4],[5]. Recently, the authors proposed a new strategy based on a modified SRP-PHAT functional that, instead of evaluating the SRP at discrete positions of a spatial grid, it is accumulated over the Generalized Cross Correlation (GCC) lag space corresponding to the volume surrounding each point of the grid [6]. The benefits of following this approach are twofold. On the one hand, it incorporates additional spatial knowledge at each point for making a better final decision. On the other hand, the proposed modification achieves the same performance as SRP-PHAT with fewer functional evaluations, relaxing the computational demand required for a practical application.

In this paper, we analyze the distribution of location estimates obtained with the modified SRP-PHAT functional with the aim of establishing a speaker detection rule to be used in a videoconferencing environment involving multiple participants. The analysis shows that location estimates follow different distributions when speakers are active, allowing to discriminate between speech and non-speech frames under a common localization framework. Moreover, the distribution of an active speaker remains almost the same for different positions inside the room, which makes easier to select a candidate location following a maximum-likelihood criterion, thus simplifying the camera-steering task.

The paper is structured as follows. Section 2 describes the conventional SRP-PHAT algorithm and our modified functional. Section 3 explains the proposed localization-based approach to speech/non-speech discrimination and speaker detection. Experiments with real-data are discussed in Section 4. Finally, the conclusions of this work are summarized in Section 5.

2. SRP-BASED SOURCE LOCALIZATION

Consider the output from microphone l, $m_l(t)$, in an M microphone system. Then, the SRP at the spatial point $\mathbf{x} = [x, y, z]$ for a time frame n of length T is defined

The Spanish Ministry of Science and Innovation supported this work under the project TEC2009-14414-C03-01.

as

$$P_n(\mathbf{x}) \equiv \int_{nT}^{(n+1)T} \left| \sum_{l=1}^{M} w_l m_l \left(t - \tau(\mathbf{x}, l)\right) \right|^2 dt, \quad (1)$$

where w_l is a weight and $\tau(\mathbf{x}, l)$ is the direct time of travel from location $\mathbf{x}$ to microphone l. DiBiase [7] showed that the SRP can be computed by summing the GCCs for all possible pairs of the set of microphones. The GCC for a microphone pair (k, l) is computed as

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \Phi_{kl}(\omega) M_k(\omega) M_l^*(\omega) e^{j\omega\tau} d\omega, \quad (2)$$

where τ is the time lag, * denotes complex conjugation, $M_l(\omega)$ is the Fourier transform of the microphone signal $m_l(t)$, and $\Phi_{kl}(\omega) = W_k(\omega) W_l^*(\omega)$ is a combined weighting function in the frequency domain. The phase transform (PHAT) [8] has been demonstrated to be a very effective GCC weighting for time delay estimation in reverberant environments:

$$\Phi_{kl}(\omega) \equiv \frac{1}{|M_k(\omega) M_l^*(\omega)|}. \quad (3)$$

Taking into account the symmetries involved in the computation of Eq.(1) and removing some fixed energy terms [7], the part of $P_n(\mathbf{x})$ that changes with $\mathbf{x}$ is isolated as

$$P'_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} R_{m_k m_l}\left(\tau_{kl}(\mathbf{x})\right), \quad (4)$$

where $\tau_{kl}(\mathbf{x})$ is the *inter-microphone time-delay function* (IMTDF). This function is very important, since it represents the theoretical direct path delay for the microphone pair (k, l) resulting from a point source located at $\mathbf{x}$. The IMTDF is mathematically expressed as

$$\tau_{kl}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{x}_k\| - \|\mathbf{x} - \mathbf{x}_l\|}{c}, \quad (5)$$

where c is the speed of sound, and $\mathbf{x}_k$ and $\mathbf{x}_l$ are the microphone locations.

The SRP-PHAT algorithm consists in evaluating the functional $P'_n(\mathbf{x})$ on a fine grid G with the aim of finding the point-source location $\mathbf{x}_s$ that provides the maximum value:

$$\hat{\mathbf{x}}_s = \arg\max_{\mathbf{x} \in G} P'_n(\mathbf{x}). \quad (6)$$

2.1. Modified SRP-PHAT Functional

Recently, the authors proposed a new strategy where, instead of evaluating the SRP functional at discrete positions of a spatial grid, it is accumulated over the GCC lag space corresponding to the volume surrounding each point of the grid as follows:

$$P''_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \sum_{\tau=L_{kl1}(\mathbf{x})}^{L_{kl2}(\mathbf{x})} R_{m_k m_l}(\tau). \quad (7)$$

The GCC accumulation limits $L_{kl1}(\mathbf{x})$ and $L_{kl2}(\mathbf{x})$ are determined by the gradient of the IMTDF corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry, as explained in [6].

3. SPEAKER DETECTION

In the next subsections, we describe how active speakers are detected in our system, which requires a previous discrimination between speech and non-speech frames based on the distribution of location estimates. To this end, we model the probability density function of the obtained locations when there are active speakers and when silence and/or noise is present.

3.1. Distribution of Location Estimates

Our first step to speaker detection is to analyze the distribution of the location estimates $\hat{\mathbf{x}_s}$ when there is an active speaker talking inside the room from a static position. In this context, six microphones were placed on the walls of the video-conferencing room and a set of 12 recordings from different speaker positions were analyzed to obtain the resulting location estimates. Figure 1(a) shows an example of three two-dimensional histograms obtained from different speaker locations. It can be observed that, since the localization algorithm is very robust, the resulting distributions when speakers are active are significantly peaky. Also, notice that the shape of the distribution is very similar in all cases but centered in the actual speaker location. As a result, we model the distribution of estimates as a bivariate Laplacian as follows:

$$p(\hat{\mathbf{x}}_s | H_s(\mathbf{x}_s)) = \frac{1}{2\sigma_x \sigma_y} \exp^{-\sqrt{2}\left(\frac{|x - x_s|}{\sigma_x} + \frac{|y - y_s|}{\sigma_y}\right)}, \quad (8)$$

where $p(\hat{\mathbf{x}}_s | H_s(\mathbf{x}_s))$ is the conditional probability density function (pdf) of the location estimates under the hypothesis $H_s(\mathbf{x}_s)$ that there is an active speaker located at $\mathbf{x}_s = [x_s, y_s]$. Note that the variances σ_x^2 and σ_y^2 may depend on the specific microphone set-up and the selected processing parameters. This dependence will be addressed in future works. On the other hand, a similar analysis was performed to study how the distribution changes when there are not active speakers, i.e. only noise frames are being processed. The resulting histogram can be observed in Figure 1(b), where it becomes apparent that the peakedness of this distribution is not as significant as the one obtained when there is an active source. Taking this into account, the distribution of non-speech frames is modeled as a bivariate Gaussian:

$$p(\hat{\mathbf{x}}_s | H_n) = \frac{1}{2\pi\sigma_{x_n}\sigma_{y_n}} \exp^{-\left(\frac{x^2}{2\sigma_{x_n}^2} + \frac{y^2}{2\sigma_{y_n}^2}\right)}, \quad (9)$$

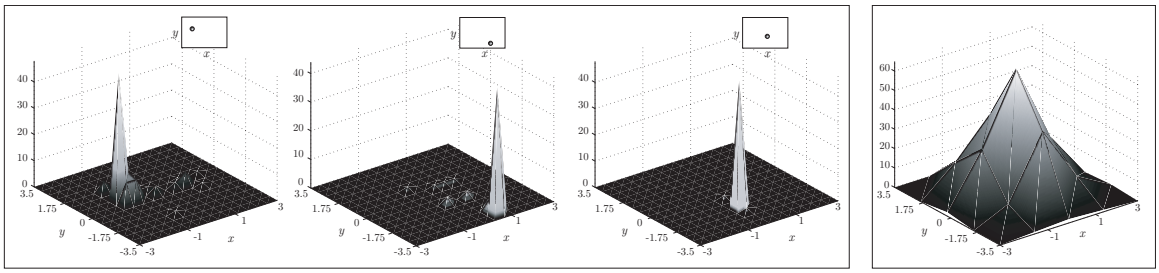


Fig. 1. Two-dimensional histograms showing the distribution of location estimates. (a) Distribution obtained for three different speaker locations. (b) Distribution for non-speech frames.

where $p(\hat{\mathbf{x}}_s|H_n)$ is the conditional pdf of the location estimates under the hypothesis H_n that there are not active speakers, and the variances $\sigma^2_{x_n}$ and $\sigma^2_{y_n}$ are those obtained with noise-only frames.

3.2. Speech/Non-Speech Discrimination

In the last subsection, it has been shown that speech frames are characterized by a bivariate Laplacian probability density function. A similar analysis of location estimates when there are not active speakers results in a more Gaussian-like distribution, which is characterized by a shape less peaky than a Laplacian distribution. This property is used in our system to discriminate between speech and non-speech frames by observing the peakedness of a set of accumulated estimates:

$$\mathbf{C} = \begin{bmatrix} \hat{x}_s(n) & \hat{y}_s(n) \\ \hat{x}_s(n-1) & \hat{y}_s(n-1) \\ \vdots & \vdots \\ \hat{x}_s(n-L-1) & \hat{y}_s(n-L-1) \end{bmatrix} = [\mathbf{c}_x\, \mathbf{c}_y], \quad (10)$$

where L is the number of the accumulated estimates in matrix $\mathbf{C}$. A peakedness criterion based on high-order statistics was evaluated. Since the kurtosis of a normal distribution equals 3, we propose the following discrimination rules for active speech frames:

$$\text{Kurt}(\mathbf{c}_x) \begin{cases} \geq 3 & \text{speech} \\ < 3 & \text{non}-\text{speech} \end{cases}, \quad (11)$$

$$\text{Kurt}(\mathbf{c}_y) \begin{cases} \geq 3 & \text{speech} \\ < 3 & \text{non}-\text{speech} \end{cases}, \quad (12)$$

where a frame is selected as speech if any of the above conditions is fulfilled.

3.3. Camera Steering

To provide a suitable camera stability, a set of target positions were pre-defined coinciding with the actual seats in the videoconferencing room. The localization system will be responsible for communicating the camera which of the target positions is currently active. This process involves two main steps. First, it is necessary to discriminate between speech and non-speech frames as explained in Section 3.2. If a burst of speech frames is detected, then the estimated target position is forwarded to the camera when it does not match the current target seat. Since all the target positions are assumed to have the same prior probability, a maximum-likelihood criterion is followed:

$$\hat{\mathbf{x}}_t = \arg\max_{\mathbf{x}_t} p\left(\hat{\mathbf{x}}_s|H(\mathbf{x}_t)\right), \quad t = 1 \ldots N_t, \quad (13)$$

where $\mathbf{x}_t$ is one of the N_t pre-defined target positions. Given that the likelihoods have the same distribution centered at different locations, the estimated target position $\hat{\mathbf{x}}_t$ is the one which is closest to the estimated location $\hat{\mathbf{x}}_s$.

4. EXPERIMENTS

To evaluate the performance of our proposed approach a set of recordings was carried out in a videoconferencing test room with dimensions 6.67 m x 5.76 m x 2.10 m. A set of 6 omnidirectional microphones were placed on the walls of the room. To be precise, 4 of the microphones were situated at the 4 corners of the ceiling of the room and the other two microphones were placed at the same height but in the middle of the longest walls. Figure 2 shows the microphone set-up, the camera location and the different seats occupied by the participants. Black dots represent the 12 pre-defined target locations used to select the active speaker seat.

The experiment consisted in recording speakers talking from the different target positions (only one speaker at each time) with the corresponding space of silence between two talking interventions. The recordings were processed with the aim of evaluating the performance of our system in discriminating speech from non-speech frames and determining the active speaker so that the camera can point at the correct seat. With this aim, the original recordings were manually labeled as speech and non-speech fragments. The processing used a sampling rate of 44.1 kHz, with time windows of 2048 samples and 50% overlap. The location estimates were calculated using the modified SRP-PHAT functional, as explained

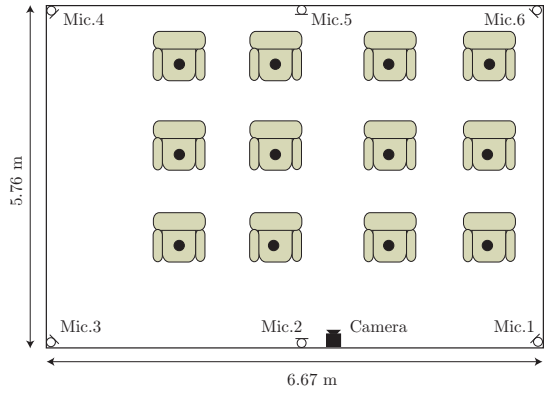


Fig. 2. Videoconferencing test room and microphones location.

Grid res.	0.5 m				0.3 m			
L	5	10	15	20	5	10	15	20
% SP	52.5	60.4	70.0	74.0	68.9	70.7	83.1	85.4
% N-SP	75.9	64.8	70.9	72.7	81.4	70.9	81.5	82.3
% T	98.2				99.6			

Table 1. Performance in Terms of Percentage of Correct Frames

in Section 2. The discrimination between speech and non-speech frames was carried out by calculating the kurtosis of the last L estimated positions, as explained in Section 3.2.

4.1. Results

Table 1 shows the percentage of correctly detected speech (% SP) and non-speech (% N-SP) frames with different number of accumulated positions $L = 5, 10, 15, 20$. Moreover, the processing was performed considering two different spatial grid sizes (0.3 m and 0.5 m). The percentage of speech frames with correct target positions (% T) is also shown in the table. It can be observed that, generally, the performance increases with a finer grid and with the number of accumulated estimates L. These results were expectable, since the involved statistics are better estimated with a higher number of location samples. Although it may seem that there are a significant number of speech frames that are not correctly discriminated, it should be noticed that this is not a problem for the correct driving of the camera, since most of them are isolated frames inside speech fragments that do not make the camera change its pointing target.

5. CONCLUSION

This paper presented a microphone array system for camera-steering to be used in a multiparticipant videoconferencing environment based on the well-known SRP-PHAT algorithm. The distribution of location estimates obtained with a modified SRP-PHAT functional was analyzed, showing that location estimates follow different distributions when speakers are active and allowing to discriminate between speech and non-speech frames under a common localization framework. The results of experiments conducted in a real room suggest that, using a moderately high number of accumulated location estimates, it is possible to discriminate with significant accuracy between speech and non-speech frames, which is sufficient to correctly detect an active speaker and make the camera point at his/her pre-defined location.

6. REFERENCES

[1] E. Ettinger and Y. Freund, "Coordinate-free calibration of an acoustically driven camera pointing system," in *Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC 2008)*, Stanford, CA, 2008.

[2] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, Washington, DC,, 1997.

[3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications*, chapter Robust Localization in Reverberant Rooms, pp. 157–180, Springer-Verlag, 2001.

[4] Hoang Do, H. F. Silverman, and Ying Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, Honolulu, HI, April 2007.

[5] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)*, New Paltz, NY, October 2007.

[6] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, 2010, accepted.

[7] J. H. DiBiase, *A high accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University, Providence, RI, May 2000.

[8] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-24, pp. 320–327, 1976.